University of Cincinnati

Date: 6/4/2015

I, Xinhua Xiao , hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Electrical Engineering.

It is entitled:

Automated Defect Recognition in Digital Radiography

Student's name: **Xinhua Xiao**

This work and its defense approved by:

Committee chair: William Wee, Ph.D.

Committee member: Raj Bhatnagar, Ph.D.

Committee member: Chia Han, Ph.D.

Committee member: Anca Ralescu, Ph.D.

Committee member: Xuefu Zhou, Ph.D.

16152

Automated Defect Recognition in Digital Radiography

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Department of Electrical Engineering and Computing Systems

of the College of Engineering and Applied Science

by

Xinhua Xiao

B.Eng. Dalian University of Technology, Dalian, China, 2009

June 2015

Committee Chair: William G. Wee, Ph.D.

Abstract

Digital radiography has been widely used for non-destructive testing in industrial production. An example is the inspection of industrial parts like turbine blades of jet engines. ADR (Assisted Defect Recognition) is an X-ray image based inspection system developed for anomaly detection of turbine blades. The system uses a reference-based method, in which statistical models are created at pixel level based on a reference model image set from good parts. And pixels of the images under test, when compared with the statistical models, which yield significant differences are called out and identified as potential defects. The system is efficient to detect low-contrast defects, but its effectiveness heavily relies on the reference model image set. When non-representative reference image sets are used, there is a high probability of false rejections. Due to variations in the production process, the reference image set may have to be adapted.

The research work proposes an automatic approach to select, based on the feature extractions of callout images of the system, a representative reference model image set for the system. Experimental results show that the proposed approach can select a model image set with a low false alarm rate and acceptable detection rate and outperforms manual approach. To adapt to the variations in the production process, an adaptive procedure based on the automatic approach is proposed to update the reference model set. Experimental results show that the proposed procedure can automatically detect significant variations and update the model set with little human intervention.

The research work also studies the impact of using reference model image sets containing images of parts with defect indications (imperfect images). A systematic procedure is proposed to evaluate the impact of imperfect images on the performance of ADR based on McNemar's test. The number of imperfect images which can be tolerated in the model set is determined for each type of defect indications.

To further improve the defect recognition rate, the research work proposes a hybrid method by combining the ADR system and a new classifier based on scan line and modified Haar-like features. An image under test is first inspected by the ADR system. If the image is not called out by ADR, scan line and modified Haar-like features are extracted at a series of specific regions. If the decision rules defining the normal

feature space which are learned from the model image set are not satisfied, the image is considered as defective. Experimental results show that the proposed method can detect all non-callout strong defective images by ADR without increasing false alarm rate. The proposed method is also applicable to the detection of strong positive images.

Acknowledgements

Foremost, I would like to thank my advisor Professor Wee for his continuous guidance and encouragement throughout my PhD study. He has taught me how to see the big picture, and think outside the box. He has helped me to skillfully handle the challenges and difficulties I have encountered in the research work.

I would also like to thank Professor Han and Professor Zhou, who have given me invaluable help and advice on research methods, paper writing, presentation skills, etc.

I want to thank my committee members, Professor Bhatnagar and Professor Ralescu for their time and thoughtful advice with regard to this dissertation.

I appreciate the help from all the former and current members of the Multimedia and Augmented Reality Lab.

I want to thank my parents, my brother and sister for their support during my time at University of Cincinnati. Thanks to my wife, Jie, for her love and accompany during the most stressful times in my PhD study.

Finally, my acknowledgement goes to GE Aviation for the financial support of my research projects.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

The highly competitive manufacturing industry has demanded higher quality and lower manufacturing costs for the past several decades. These requirements have led to great technological advances of automation in manufacturing processes [1]. One of the critical components of any manufacturing process is part inspection. Part inspection consists of tasks of measuring varied attributes of the parts, such as dimensions, shape, mass, locations and sizes of machining operations, to ensure that they meet required quality standards [2]. Part inspection usually employs methods of Non-destructive Testing (NDT) in order not to induce damage to the inspected parts and affect their future usefulness. NDT methods include diverse techniques, like radiographic X-ray imaging, fluorescent penetrant inspection, and eddy-current testing. Among these techniques, radiographic X-ray imaging is the most commonly used for locating abnormal features that are located inside the manufactured parts, e.g., the aluminum wheels, steering gears of cars, and the turbine blades of jet engines [3,4].

A variety of methods have been developed for automated anomaly detection of industrial parts via computer-aided analysis of the X-ray images [5]. These methods can be divided into two categories: reference- and non-reference-based methods [6]. The methods in the latter category, the non-reference-based methods, are often used when the reference images are unavailable [6]. Various kinds of defects or anomalies are defined, and methods such as expert systems, artificial neural networks, or general filters are used to differentiate them from the characteristics of the normal images [7-11]. Due to the difficulty of defining all possible defects or anomalies, the application of these methods is quite limited. When reference images are available, methods in the first category, the reference-based methods are usually utilized since the reference images or their statistics can be chosen as the benchmark. A test image is compared with the benchmark, and if significant differences exist, then the test image is classified as anomalous [4-5, 12-14].

The reference-based methods allow a set-actual comparison which is not possible with the non-reference

based methods, and are efficient for detecting low contrast defects [16]. However, the performance of the reference-based methods relies on the reference images selected from good parts [15, 16]. Besides, parts can vary within the specification during the production process [16]. For example, for the aluminum die casting, abrasions and wear are common during the lifetime of a mold, and also sand cleaning of the molds leads to variations in wall thickness. These subtle variations are visible in the X-ray images. This makes difficult the comparison of older reference image sets with current images under inspection.

Clearly, there is an industry-wide demand that an automatic approach be developed to find a representative reference data set to create models and an adaptive method be found to update the reference image data if high parts' variation occurs. Besides, the reference data set are selected from good parts. It is labor intensive and subjective for human experts to classify part images into good and bad ones. To save labor cost, it is natural to ask whether the reference data set can tolerate defective images or not. In such a situation, it is necessary to investigate methods of evaluating performance difference for model sets with and without defective images. In addition, for any automated anomaly detection method, the performance metrics including detection rate and the false rejection rate can hardly be ideal. There is always a need to develop new methods to further improve the detection.

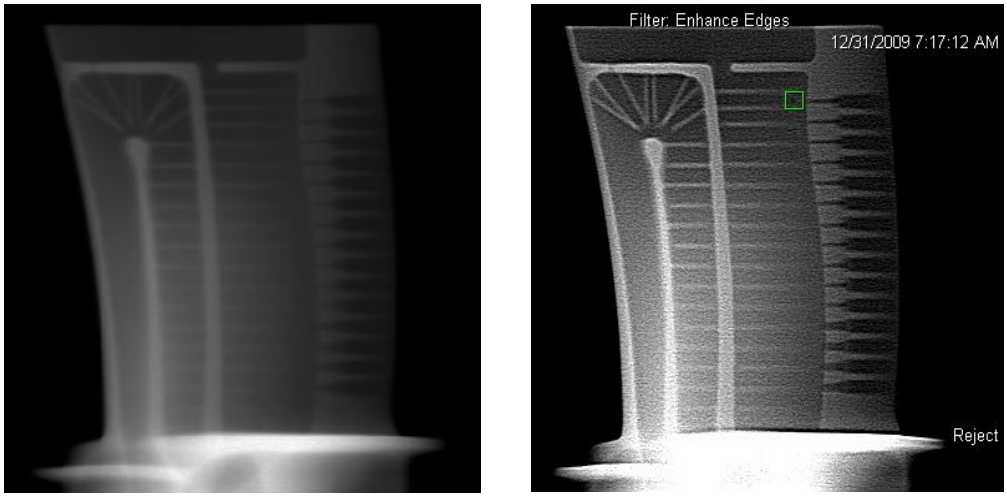
1.2 Research Scope

The research work in this dissertation focuses on improvement in a typical reference-based part inspection system, the assisted defect recognition (ADR) for turbine blades of jet engines [4, 15, and 47]. The ultimate goal is to develop systematic and applicable methodologies in the following four major research areas:

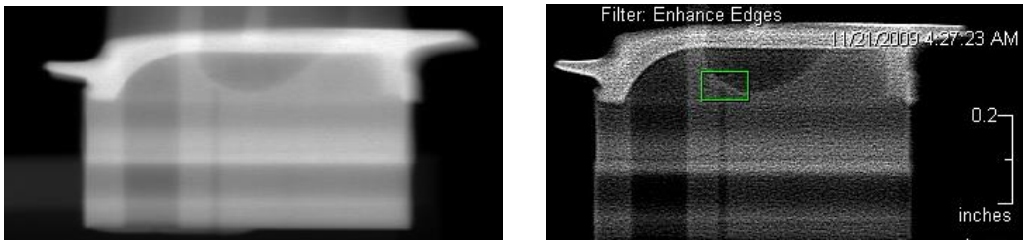
- Find a representative reference model image set from a large set of anomaly free images for the ADR system;
- Adapt the selection of the reference image set to normal parts' variation so that a low false rejection rate is ensured from time to time;
- Evaluate the impact of defective images if included in the reference image set on the performance of ADR;
- Further improve the recognition rate of the current ADR inspection system.

1.3 Challenges

Part defects manifest themselves in the image by some change of brightness and may have an arbitrary shape and size [48]. Defects of a turbine blade, which is a basic component of a gas turbine and is responsible for extracting energy from the high temperature, high pressure gas produced by the combustor [54], correspond to changes of brightness in the X-ray images and can be as small as 10-12 pixels in size, and have very low contrast, and some may hide behind the structure of the part. Figures 1.1 (a) and (c) show the original images for a part with a small defect indication. Figures 1.1 (b) and (d) are the corresponding enhanced images with the defect labeled by a green bounding box. As seen from Figure 1.1, the defects are very small and with very low contrast, making it difficult to detect.



(a) Original Trailing view image with a defect (b) Enhanced image of (a) with the defect labeled



(c) Original Root view image with a defect (d) Enhanced image of (c) with the defect labeled

Figure 1.1. The original and enhanced image of a turbine blade with a small defect indication labeled with a green bounding box.

On the other hand, changes of brightness in the image does not necessarily caused by defects, possible other sources [4] include

- image misalignment, which are related to part placement, pose changes and distortion in the imaging process;
- variations in imaging source and detector;
- part-to-part variations within normal specification range.

How to rule out these irrelevant changes in the image and identify the truly defect indications with a high reliability is a big challenge.

Besides, the number of parts, especially the number of turbine blades, could be large in a given production line, and the specifications very stringent, thus the industry requires the inspection be fast and with high accuracy.

1.4 Contributions

The research work focuses on solving practical and important industrial application problems, more specifically, it theoretically and practically contributes in the following:

- A feature-based approach, called ADR Model Optimizer, is proposed to automatically select a representative model set from a large defect-free image set, which is applicable for different types of blades and different views of the blade, and outperforms than manual approach with less computation time and lower false alarm rate.
- An adaptive model selection procedure is proposed to adapt the reference image data to the parts' variation detected during the production process, which involves little human intervention and can make sure that the reference image set represents the current state of production process.
- A systematic methodology is put forward to evaluate the effect of including images of parts with anomalies into the reference data on the performances of ADR and determines the number of defective images allowed in the reference data. Four methods are put forward to include defective images into the reference data based on the defect type, location, and defect area size. Different statistical tests are investigated to determine if there is significant difference for the ADR inspection system before and after the including.
- A new classifier based on scan-line and modified Haar-like features is proposed to further improve the detection of ADR. The new classifier is efficient for implementation and can identify defective images which are not called out by ADR without increasing the false alarm rate.

1.5 Dissertation Structure

The rest of the dissertation is organized as follows.

Chapter 2 is the literature review. Image-based automated defect recognition methods are reviewed. The state of the art of commonly used statistical tests for classifier performance is presented.

In Chapter 3, the reference model image set problem is investigated. An automatic approach based on the features of the output images of ADR is proposed to select a reference model set. Various experiments are implemented to verify the feasibility of the proposed approach.

Chapter 4 addresses the adaptive model selection problem. A systematic procedure is put forward and illustrated in details as to how to detect the parts' variation during the production process and to update the reference model set.

In Chapter 5, the problem of evaluating the impact of images of parts with anomalies (imperfect images) in the reference model set on the performance of ADR is studied. Four methods are put forward as how to include the imperfect images into the reference set to produce the worst effect on the performance of ADR. McNemar's test is presented and used for evaluating the effect of imperfect images.

Chapter 6 investigates how to detect defective images which cannot be detected by ADR. A new classifier based on scan line features and modified Haar-like features is presented for improve detection.

Chapter 7 is the conclusion part.

2 Literature Review

This section reviews the state-of-the-art of image-based defect recognition methods and statistical tests for classifier performance evaluation.

2.1 Automated Defect Recognition Based on Radiographies

Radiography is widely used for non-destructive testing in industrial production for casting part inspection [5], as it can locate casting defects inside the testing parts which are not visible to the naked eye [5]. Based on radiographies, there are many methods developed for defect recognition in industrial applications. These methods can be divided into two categories: 1) reference-based methods; and 2) non-reference based methods [6, 18].

2.1.1 Reference-based Methods

The reference-based methods use reference based information – like the reference image data or the created models based on the reference image data as the benchmark. If there is a significant difference between the tested image and the benchmark [6, 18], then corresponding inspected part of the tested image is classified as defective. The reference-based methods allow a set-actual comparison which is not possible for other methods, and particularly apply to low contrast defects [19]. The PXV (Philips X-ray Vision), the AI (Automatic Inspector by YXLON), and the ADR (Assisted Defect Recognition by Generic Electric) are well-known systems for industrial applications today using the reference-based methods for defect recognition [4, 6, 5, 23-25].

2.1.1.1 The PXV and AI Radioscopic Test System

The PXV was developed by Philips Industrial X-ray GmbH as a fully automatic radioscopic testing device [5, 26]. The system was further developed by YXLON International X-ray GmbH, and is called AI (Automatic Inspector) [5, 27]. The PXV and the AI follows the following main strategy for defect detection of aluminum die cast pieces [5, 23- 27]:

- Create a defect-free reference image from the original image of the test part by filtering;
- Obtain a difference image by subtracting the processed reference image from the original image;
- Build a binary image marking defect indications from the difference image by thresholding.

Figure 2.1 shows the defect recognition process of the above strategy [24]. The AI uses a Neural Network filtering method based on a Hopfield-Tank NN to process the original test image to obtain the reference image [23, 27]. The PXV uses several complex filters for specific regions of interest of the original image to obtain the reference image [23, 25]. Both methods adopt a supervised training, and during the training human experts have to stay in place at the batch testing system and should have a deep understanding of all chosen parameters for the filters [5, 23-27]. For both methods, defects bigger than the filter kernel cannot be detected.

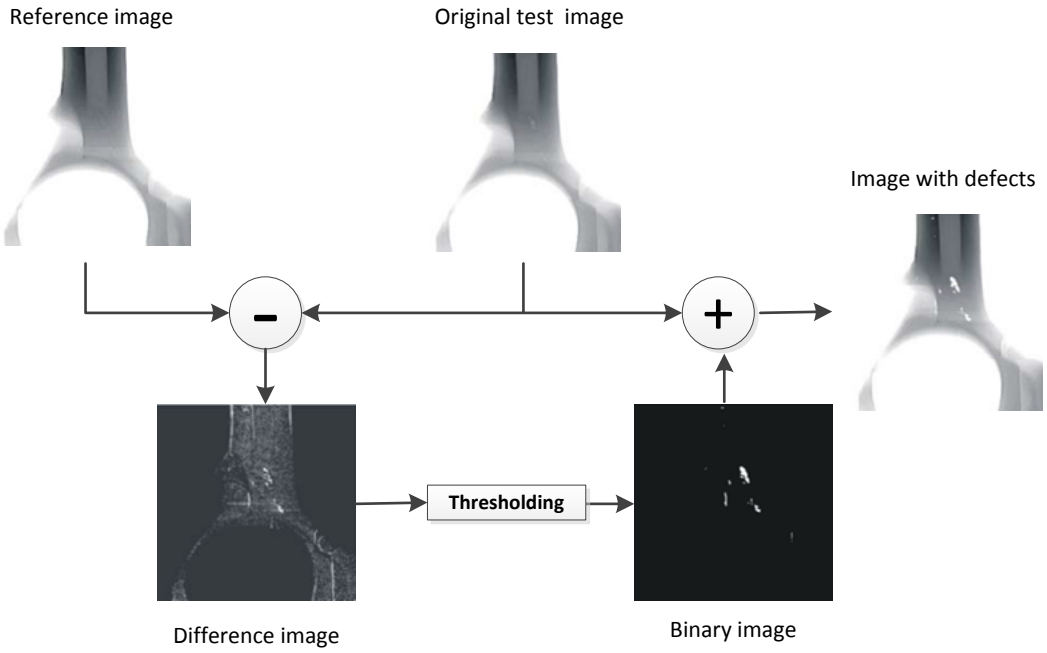


Figure 2.1. The defect recognition process for the PXV and the AI. (Source: [24])

2.1.1.2 The ADR Inspection System

Compared with PXV and AI, the ADR inspection system developed by GE Aviation does not create the reference image from the test image, but select a set of reference images, and build a statistical model at each pixel based on the reference image set. For the test image, any pixel not fitting the model would be

considered as a potential defective pixel [4, 6]. For a turbine blade, X-ray images are taken at multiple views with 14 bits depth in grayscale. Figure 2.2 shows the process of X-ray image acquisition process for turbine blade type "A" with four different views.

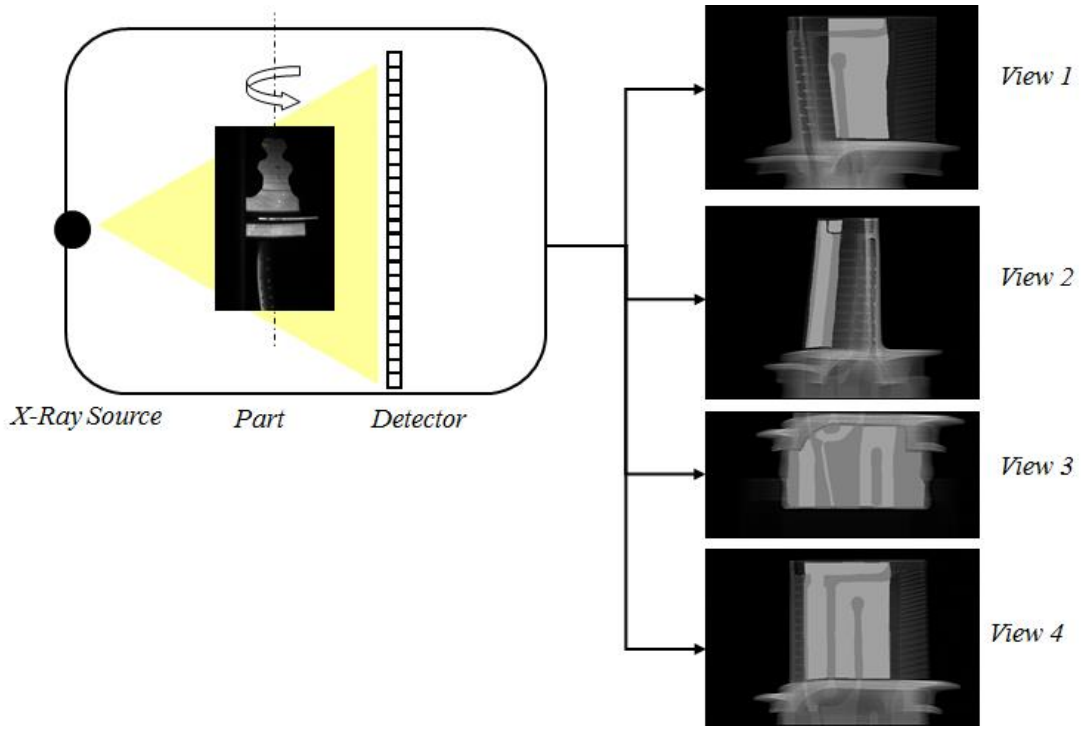


Figure 2.2. X-ray image acquisition for turbine blade type "A" with four different views.

The ADR uses each view separately to inspect the blade. Figure 2.3 shows the diagram of ADR system defect recognition process. ADR is an X-ray inspection system used for anomaly detection of turbine blades of jet engines.

The system consists of two phases: a modeling phase and an evaluation phase [4]. In the modeling phase, a set of reference image set are first aligned by registering to a template image [49], a sample image before and after registration is shown in Figure 2.4. From Figure 2.4, we see the image before registration is slightly inclined to the left, and after registration the image is upright. The aligned image is normalized by using a spatially varying median filter to remove fine structures and details for qualitative evaluation [4]. Nonparametric Parzen-window's approach is then used to build a statistical model at each pixel based on

the low-level features extracted from a set of aligned and normalized reference model images from defect-free blades. The model is defined as

$$\mathcal{M} = \left(p_j(u,v), p_j^{\alpha}(u,v), I_T(u,v), I_0(u,v), Q_j, S_j, \sigma_j\right) \tag{2.1}$$

where $p_j(u,v)$ is the defect probability at pixel(u,v), $p_j^{\alpha}(u,v)$ is the defect prior at pixel (u,v) based on domain knowledge, I_T is a template image chosen from the good parts and used for spatial alignment, $I_0(u,v)$ is the baseline image for appearance normalization, j is the defect index,

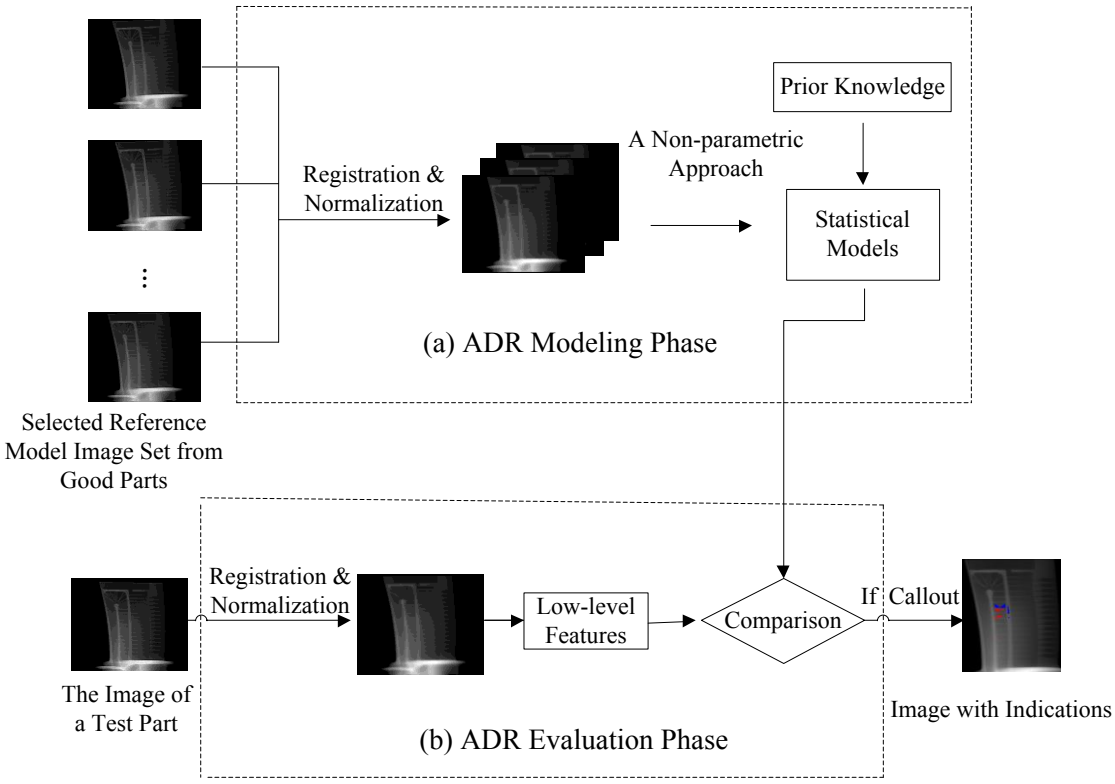


Figure 2.3. Diagram of defect recognition process for the ADR.

Q_j is the probability threshold separating normal from abnormal variations, S_j is the minimum defect size, and σ_j is the standard deviation of the Gaussian kernel [4]. Low-level image features like the intensity value are extracted, and a non-parametric statistical distribution $p_j(u,v)$ is created for each feature at pixel(u,v). For the pixels with probabilities over the threshold Q_j, 8-connected component algorithm is used, and only

regions larger than size S_j are assumed as defects [4].

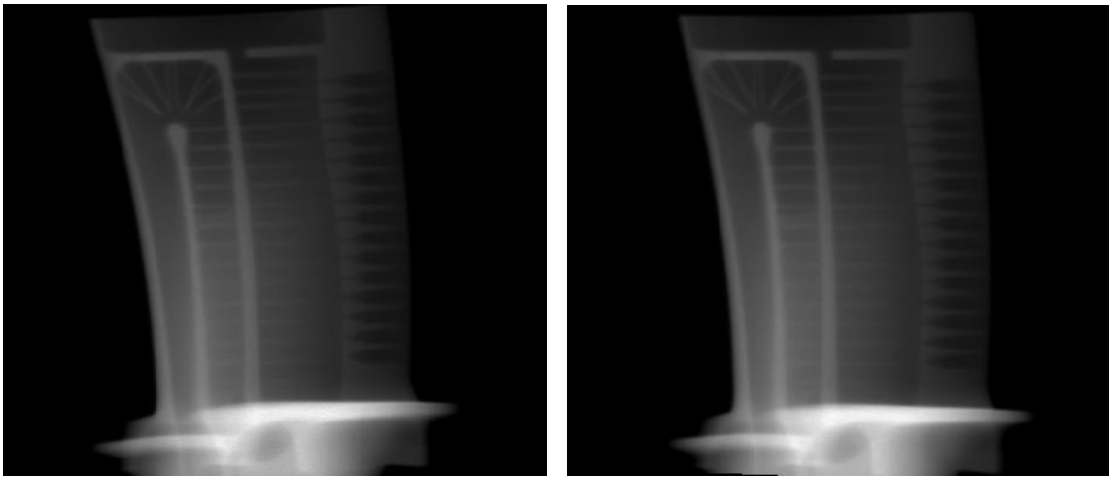
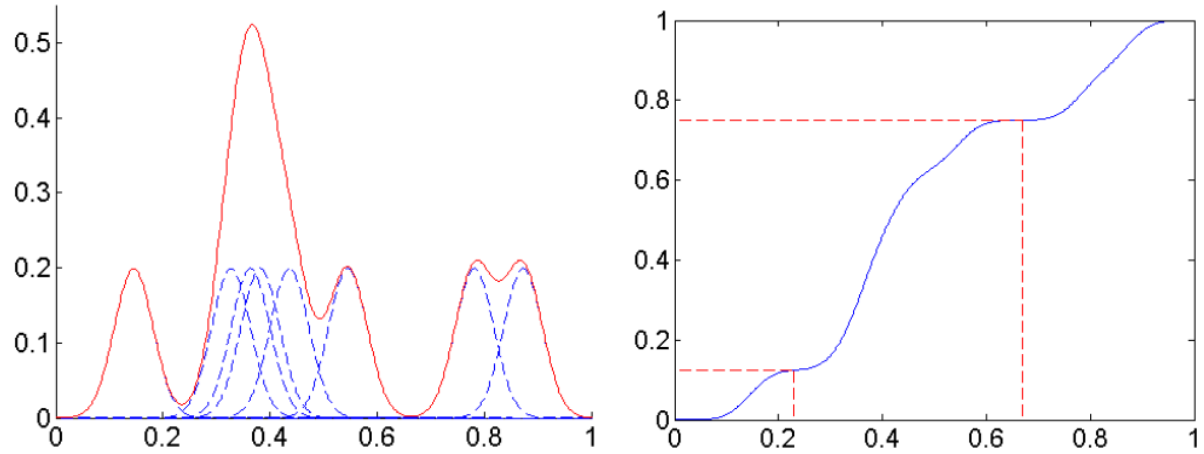


Figure 2.4. A sample image before and after registration.

As an example shown in Figure 2.5 (a), intensity values x_n at pixel (u, v) are collected and normalized from n images with $n = 8$, and a Gaussian kernel is fit at each value as shown the dashed blue lines [4]. The summation of the Gaussian mixture results in the PDF in solid red line. The CDF is computed as shown in Figure 2.5 (b), and normal range [0.22, 0.68] is determined. Anything out of this range is considered containing potential positive or negative material.

In the evaluation phase, a test image is preprocessed by the same operations including registration and normalization as in the modeling phase. Low-level features of the preprocessed test image are then extracted, and the probability of each pixel being normal or abnormal is calculated, and compared with the built statistical models. Pixels are called out if the probability is over threshold Q_j and the defect area size is larger than S_j, and are marked as blue and red in the output image, representing less material and excess material respectively. The ADR can detect defects as small as 10-12 pixels in size. The performance of ADR relies heavily on the reference model images.



(a) (b)

Figure 2.5. (a) Probability density function as a Gaussian mixture by Parzen window density approximation. (b) Cumulative density function for the probability density function. (Source: [4])

2.1.2 Non-reference based Methods

The non-reference based methods typically use image processing tools (for example, field flattening to enhance the contrast, image sharpening to make the defects obvious) [18, 20], expert systems [11], artificial neural networks [9, 21] and fuzzy logic [22]. These methods can detect defects without prior knowledge of the part structure. The prerequisite for these methods is the existence of common properties which consistently define all kinds of defects and differentiate them from the non-defective images [1, 3, and 5].

The ISAR (Intelligent System for Automatic Roentgen-inspection) is a non-reference based method developed by the Fraunhofer Institute for Integrated Circuits [5, 23, and 28]. The system identifies the die cast pieces, and performs an examination specifically for that piece [5]. After the die cast piece is identified, X-ray parameters, testing criteria, translocation of the handling device and inspection positions are selected by the user [5]. The inspection is performed with the aid of an edge preserving median filter, a so-called COMMED-filter (COMbined Median) [5, 28]. The filter performs a local image restoration on the original X-ray image without a prior knowledge of the test piece structure [23]. The ISAR can differentiate the structure of the test piece (edges, corners, bore holes etc.) from structures not part of the piece [5]. The

ISAR does not use prior-knowledge of the test pieces, and thus is difficult to differentiate between noise, structure and defects [23].

2.2 Statistical Test for Classifier Performance Comparison

Over the recent decades, there has been an increasing awareness of the need of statistical analysis to conduct the performance comparisons of different learning algorithms or classification models in the machine learning area [29-33]. Statistics allows us to decide based on the experimental results which learning algorithm or classification model outperform others [50, 51, and 52].

Statistical tests for statistical analyses can be divided into two categories: parametric and non-parametric, depending on the concrete type of data employed [50]. Parametric tests are usually based on the assumptions of independence, normality, homoscedasticity [55, 56]. Non-parametric tests are often used when these assumptions are not met.

The widely used parametric test to determine significant differences between two learning algorithms or classifiers is the paired t-test [55]. As a parametric test, the paired t-test requires conditions of independence, normality and homoscedasticity [56], which might not the case in most experiments in machine learning and pattern recognition [31, 58]. The alternative nonparametric test of paired t-test is the Wilcoxon signed-ranks test [59], which is less-powerful than t-test but more safe for use.

To compare classification rates (sensitivity, specificity) among multiple predicative models, McNemar's test is often used [59], e.g., predicating prostate cancer from diagnostics tests and patient characteristics [29, 59-62].

To address classifier accuracy comparison, a 5x2 cv t test is proposed to decide which of the two algorithms under study will outperforms the other on a given test data set [30]. Five approximate statistical tests are examined and compared experimentally to determine their probability of incorrectly detecting a difference when no difference exist (type I error), including the McNemar's test, z test for the difference of two proportions, resample paired t test, k-fold cross-validated paired t test and the proposed 5x2 cv test

[29, 30]. Dietterich in [30] finds for algorithms that can be executed only once McNemar's test is the only test with acceptable type I error, and for algorithms that can executed ten or more times, the 5x2 cv test (5 replications of 2-fold cross-validation) is slightly more powerful than McNemar's test.

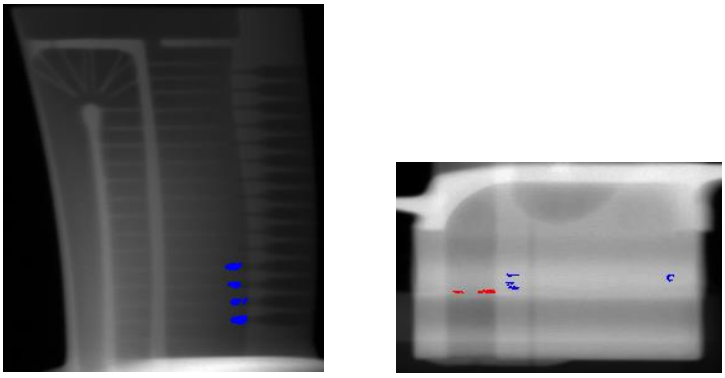
Bouckaert and Frank [32] argue the replicability of a test. They compare empirical measures of replicability with the performance of 5x2 cv test, and find that the 5x2 cv test dissatisfactory and opted for the corrected resample t-test [29, 32].

Bradley [33] investigates the use of the area under the receiver operating characteristic (ROC) curve (AUC) as a performance measure for classifiers including C4.5, K-Nearest Neighbors, etc. Bradley finds that AUC exhibits many desirable properties: increased sensitivity in Analysis of Variance (ANOVA) tests; decision threshold independent; and invariance to a prior class probability.

3 Research Problem I: Reference Data Selection for the Assisted Defect Recognition System

3.1 Overview of the Assisted Defect Recognition System

As illustrated in Section 2, ADR is an X-ray inspection system used for anomaly detection of turbine blades of jet engines. As shown in Figure 2.3, the system consists of a modeling phase and an evaluation phase. In the modeling phase, a non-parametric approach following the Bayes rule is used to build a statistical model at each pixel based on the low-level features extracted from a set of preprocessed model images of anomaly-free blades [4, 15]. The preprocessing operations of the images include registration and normalization: images are spatially aligned with a template image by pairwise image registration, and then normalized by using a spatially varying median filter. The low-level features extracted from the model images can be the intensity or texture. For simplicity, the image intensity is used as the low-level feature. In the evaluation phase, a test image is registered with the template image, and normalized by using the same filtering method in the modeling phase. Low-level features of the preprocessed test image are then extracted, and compared with the built statistical models. The pixels that do not fit the models are marked as blue and red in the output image, representing less material and excess material respectively, as illustrated in Figure 3.1.



(a) Trailing Edge View (b) Root View

Figure 3.1. Callout images of ADR with indications for blade type A, blue indication represents less material, red represents excess material.

Like any other recognition system, ADR is not ideal to detect all the anomalies of the parts under inspection, and in some cases, ADR makes misclassifications between the normal and anomalous images. If a normal test image is wrongly classified by ADR as an anomalous image, it is called a false alarm, and the corresponding callout image is called a false positive image. Accordingly, false alarm rate (FAR) is defined as:

$$FAR = \frac{n_{fp}}{N_{nor}} \tag{3.1}$$

where n_{fp} is the number of false positive images, and N_{nor} is the number of the normal test images. If an anomalous test image is correctly identified by ADR and called out, it is called detection, and the corresponding callout image is called a true positive image. The detection rate (DR) is defined as:

$$DR = \frac{n_{tp}}{N_{anor}} \tag{3.2}$$

where n_{tp} is the number of true positive images, and N_{anor} is the number of the anomalous images. FAR and DR are the performance metrics of ADR. Ideally, FAR is zero, and DR is one, i.e., there are no false alarms and no missed anomalies.

3.2 Problem Statement

ADR runs anomaly detection of turbine blades by comparing the corresponding images of the blades to the statistical models, generated from a defect-free image set, the reference model set. The performance of ADR relies on the reference model set, as shown in Figure 3.2. Currently, human experts select the model set via analysis of a large image set. The large image set is classified into two exclusive subsets: the normal set, denoted by T, and the anomalous image set, denoted by D. The reference set are drawn from T. The human expert performs the selection of the model set by guessing and iterative testing, which is subjective and time consuming. It is clear that automating the selection process can be quite meaningful [15]. The

model selection problem is formulated to develop an automatic method for determining a model set from a normal image set T, to satisfy the following conditions:

1) The model set size is as small as possible;

2) FAR is low enough while DR is acceptable to meet the industrial requirements;

3) The selection method is applicable to different blade types and varied views of the blade;

4) The model set can be updated when becoming inaccurate for anomaly detection for newly generated images.

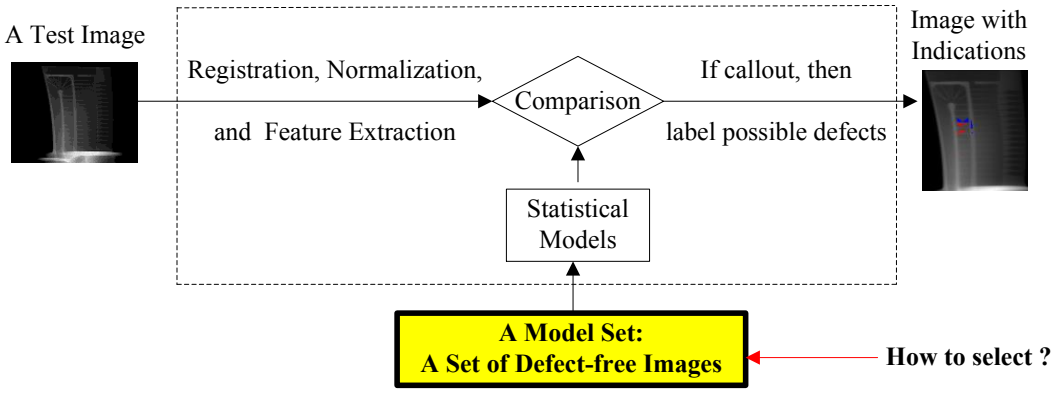


Figure 3.2. The performance of ADR relies heavily on the model set.

3.3 Proposed Model Set Selection Approach – ADR Model Optimizer

The selection of the model image set is a decision making problem: a decision making method should be devised to produces a choice of a model set which ensures a high performance of ADR [15]. The decision making process is based on features extracted from X-ray images of the blades. To be specific, as illustrated in Figure 3.3, given a large size of anomaly-free images (T), e.g., tens of thousands, the main task is to develop an approach to find a model image set by using feature extraction to satisfy the performance need of ADR: FAR of ADR is low with DR being high to meet the industrial requirement.

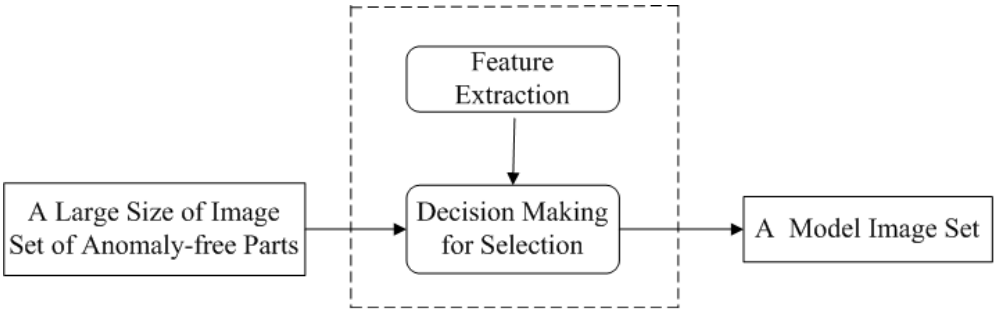


Figure 3.3. Framework of the model selection problem of ADR.

Features are measures or descriptors used to represent the original raw data in a lower dimensional space upon which applications like object classification and decision-making are performed. Desirable features are essential to any classification or decision making problem. To find a set of desirable features is usually an extremely difficult task and very much problem dependent [37]. For the model selection problem, features to be extracted are based on the analysis of the statistical approach ADR uses for anomaly detection and observations of the callout images. Features used are extracted from the callout images, and include indication type and indication size. The indication types consist of less material and excess material, and are represented by corresponding pseudo colors by comparing to the built statistical models. Two colors, as shown in Figure 3.1, blue and red, are used to indicate whether or not the intensity is in the normal range of the statistical model at each pixel. The number of the indication pixels represents the indication size. A larger size of the indications in a callout image usually represents more anomalies exist with respect to the statistical models. The usage of the extracted features to select a model set is stated in the following part.

Generally, the approach is to select a model image set based on the indication features in a two-step way: in the in the first step, compared with the statistical models built on a random initial model set, false alarm images from T are called out with indications, and we select a set of images with their corresponding callout images having the largest indication size for each indication type; in the second step, using the similar methodology, additional images are to be selected from the initial model images.

The rationale to select images featuring the largest indication size for either indication type is: intuitively,

the total indication size of the false positive image set (f_p) is usually in directly proportional to the size of f_p , which determines FAR. The inclusion of images with largest indication sizes into the model set, would reduce the total amount of indications, and thus decrease FAR. The reason for there are two steps in our approach is: a random initial model set is used, and if tested in the evaluation phase, the images in the initial model set will not be called out and marked with indications by ADR since the statistical models are built from them as the benchmark for comparison. If there is only one step, images in the initial model set would not be in the selection candidates even some of them might be very representative. In the following part, the two-step selection approach, called ADR Model Optimizer, is explained in details.

3.4 The Procedure of the Proposed Approach

This section gives the detailed processing procedure to implement the two-step selection process [15]. Given an anomaly-free image set, T, with the size denoted by n(T), the approach is to select a model image set, M12, from T in two steps.

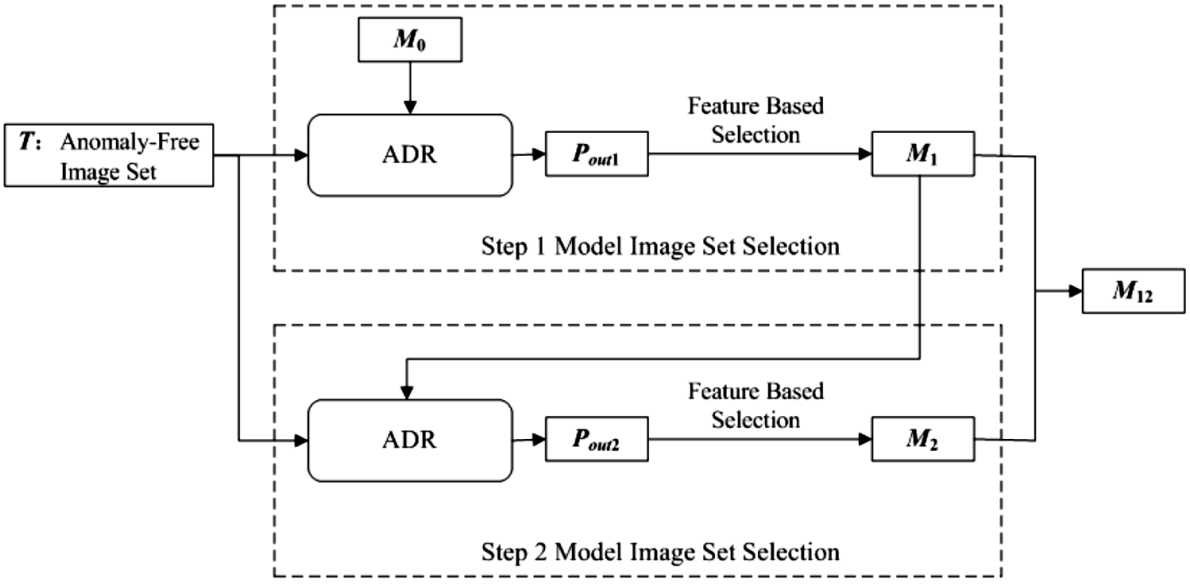


Figure 3.4. The procedure of the proposed ADR Model Optimizer.

As shown in Figure 3.4, given an anomaly-free image set T, the approach selects a model image set M12 from T in two steps, where M12 = M1 ∪ M2. M1 and M2 are two model image sets to be selected in Step 1 and 2, respectively, with the corresponding sizes n (M1), n (M2) ≪ the size of set n (T).

Step1 - The selection of the model set, M1, as shown in Figure 3.4:

1) Randomly select a set of model images, M0, where M0 ⊂ T

2) Feed M0 into ADR as the model set, and use T as the test set.

3) Run ADR to inspect every image in T. The images considered as defective images by ADR would be called out with indications marked in the callout images. Define the callout image set as Pout1.

4) Based on the indication features of Pout1, including the indication size, types and locations, select M1 from Pout1.

Step 2 - The selection of the model set, M2, as shown in Figure 3.4:

Replace the model set M0 with M1. Repeat 2), 3), 4) in Step 1. Note the callout image set in this step is defined as Pout2, and based on the features of Pout2, M2 is selected from Pout2.

The final selected model image set is M12, where M12 = M1 U M2.

3.5 Experiments Results and Discussion

This section presents the experiments to verify that the proposed approach resolves the model-image selection problem. That is, the proposed approach should select a model set, M12, to meet the requirements, including a low FAR with an acceptable DR, n(M12) being small, applicable to different types of blades and varied blade views, and updating M12 as new batches of images are generated and the model becomes inaccurate for anomaly detection for the new images.

Figure 3.5 shows the experimental result to verify the general effectiveness of the proposed approach. The blade type used here is “A", and the blade view is View 2, with Nnor = 1940. Set n(M0) = 130, n(M1) = 100, and n(M2) = 30. FAR is normalized to 100% at the starting point. From the figure, we see, compared to the starting point, FAR decreases by 70% in Step 1, and by nearly 90% in Step 2 with the value to be

10%. The blue line is the result of the manual selection approach. Compared with the manual selection approach, the value of FAR based on M12 is about 30% lower. For the metric of DR, both the manual approach and proposed approach have the similar performance, which will be discussed in Figure 3.8. As to computation time, the proposed approach is much faster than manual approach. The proposed approach needs to test the normal 1940 images three times using ADR with the initial model set M0, the M1 model set and the final set M12, and the total computation time is about 9-16 hours. However, the manual approach is based on guessing and iterative testing. It takes at least one to two weeks. From the above, we see the proposed approach has better performance and less computation time than the manual approach.

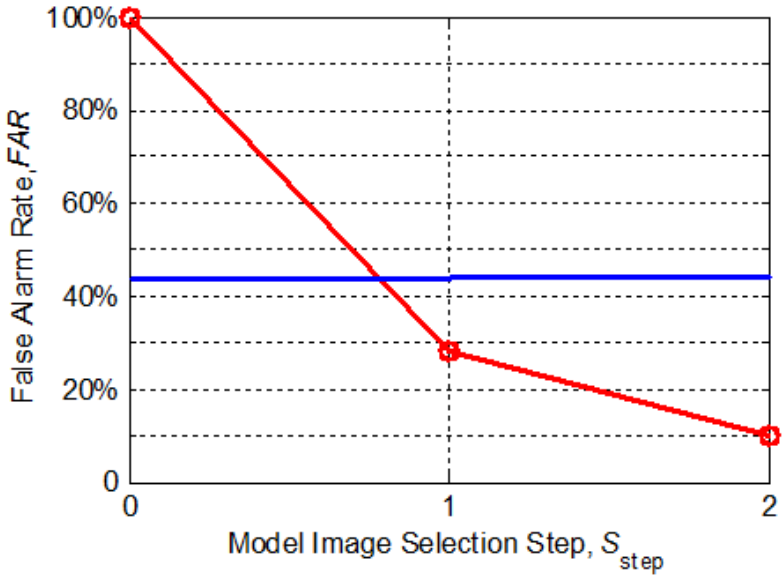


Figure 3.5. Results of the model selection approach for View 2 of blade type “A".

Note: FAR is normalized with respect to (w.r.t.) the starting point.

Figure 3.6 and 3.7 show that the experimental results by applying the two-step selection approach to different blade types and varied views of the blade. From Figure 3.6, we see for the same view (View 1), FAR for blade type “A" and “B" is decreased significantly in two steps, indicating the approach is applicable to different blade types for the model selection. As shown in Figure 3.7, for the same blade type (“A"), FAR for all the four views is greatly reduced in two steps. This result demonstrates the approach is

applicable to different views for the model selection.

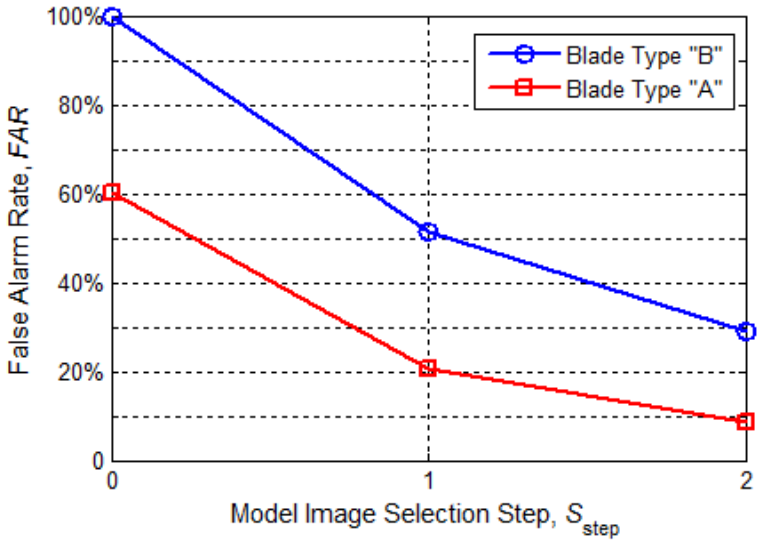


Figure 3.6. FAR against Sstep for View 1 of blade type “A" and “B".

Note: FAR is normalized w.r.t. the starting point for the blade type “B”.

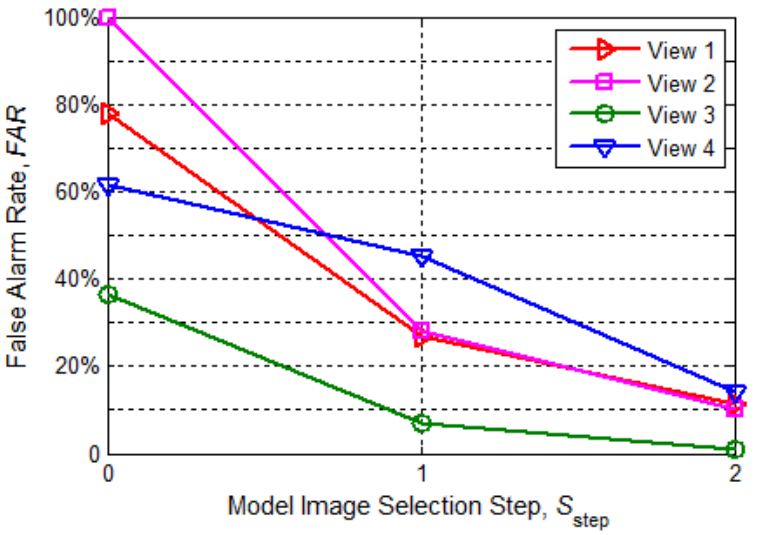


Figure 3.7. FAR against Sstep for all the fours views of blade type “A".

Note: FAR is normalized w.r.t. the starting point of View 2.

Figure 3.8 shows the study of the impact of model size, n (M12) on FAR and DR. The blade image used here is View 1 from blade type “B". The total number of the anomaly-free images is 4032, that is, Nnor =

4032. The truly anomalous image number Nanor = 70. The model image set, M12, should be selected from the 4032 anomaly-free images. Set the model size, n (M12), to be 50, 70, 90, 110, 130, 190, 240 and 300. Shown from Figure 10, as n (M12) increases from 50 to 130, the value of FAR decreases fast, and remains stable when n (M12) > 130. However, the value of DR changes slightly with the increase of n (M12), and is similar with the manual approach (the green line). We note that due to the computation complexity, n (M12) should be as small as possible. To ensure a low value of FAR with an acceptable value of DR, and consider the computation time, we should pick an optimal value of n (M12). It is noted that n (M12) $\in$ [110; 130] is the optimal value interval.

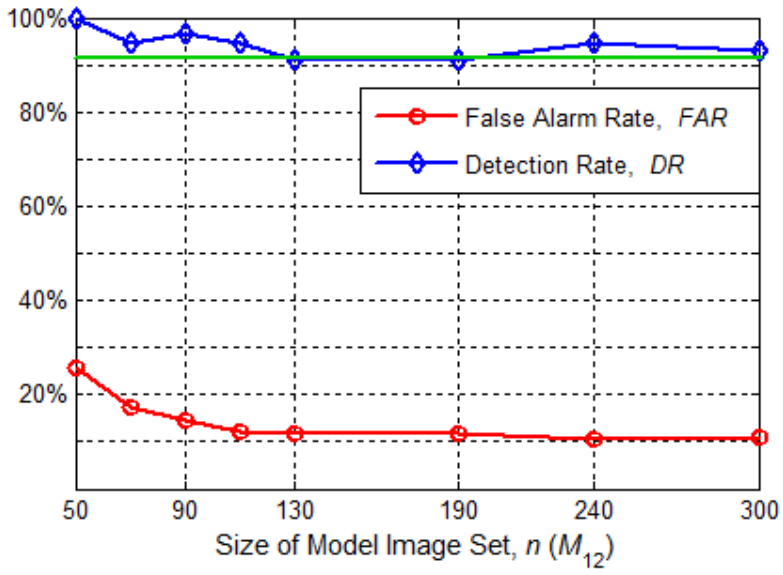
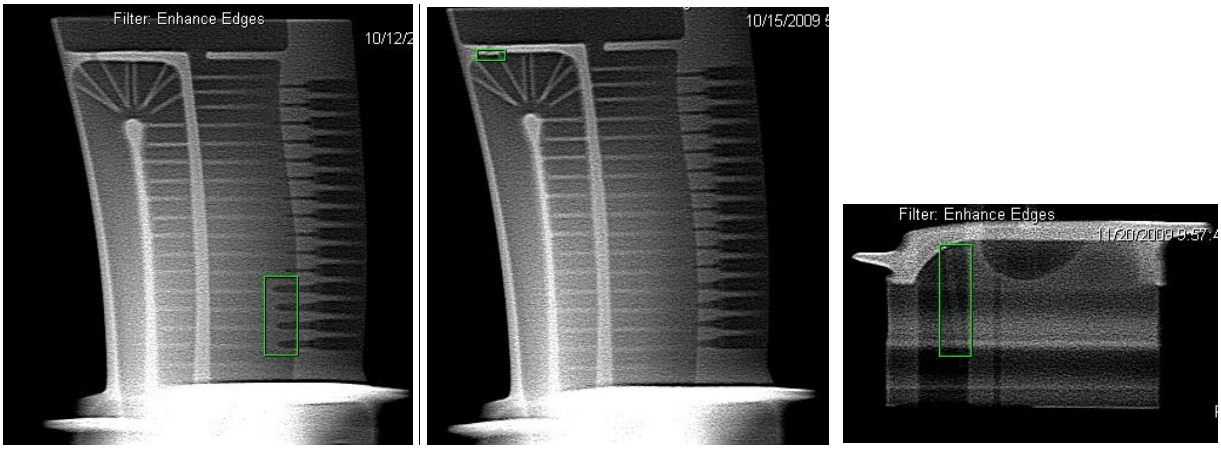


Figure 3.8. FAR and DR against n (M12) for View 1 of blade type "B".

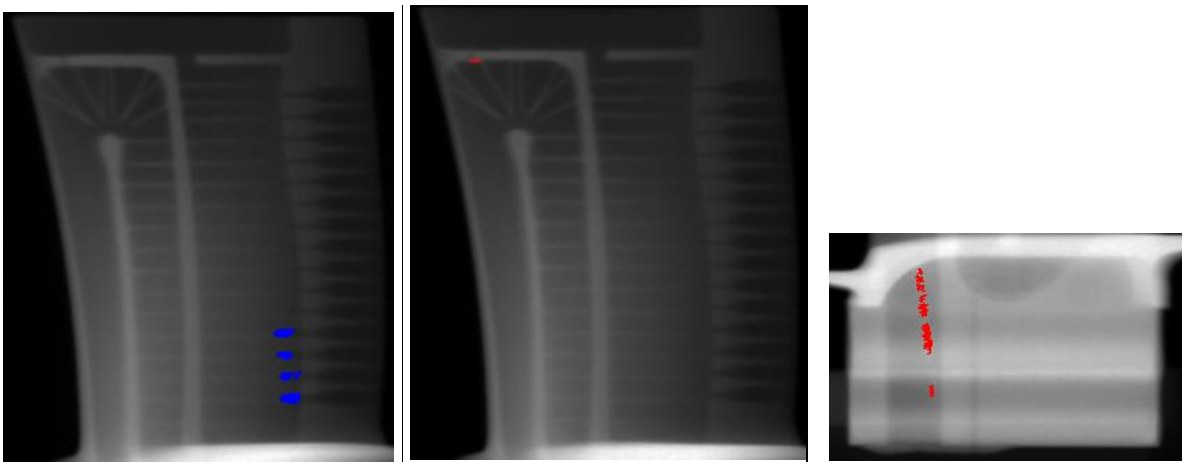
Note: FAR and DR are normalized w.r.t. the value of DR at n (M12) = 50.

From the experiments, we see the proposed approach can select a good model set from a large size of defect-free images with low false alarm rate and acceptable detection rate. We further investigate the qualitative results of detection of defective blades using ADR based the selected good model set. Figure 3.9 shows the ground truths of defect indications labeled with green bounding boxes in the images by human experts for blade type "B", and Figure 3.10 shows the corresponding detection results using ADR based on a selected good model set. For a better visualization, the human labeled images are enhanced by filtering.



(a) View 1 with negative defects (b) View 1 with postive defects (c) View 3 with positive defects

Figure 3.9. Ground truth of defect indications labeled with green bounding boxes in the images by human experts for blade type "B".



(a) View 1 with negative indications (b) View 1 with postive indications (c) View 3 with positive indications

Figure 3.10. The corresponding detection results using ADR based on a selected good model set.

From Figure 3.10, we see the defect indications marked by ADR are relatively accurate in the locations corresponding defective blades based on the selected model set. This demonstrates the correctness of the ADR and the selected model set.

3.6 Conclusions

An automatic selection approach is proposed to resolve the model image selection problem of ADR. The approach selects model images based on the features extracted from the callout images of ADR. Experimental results demonstrate that the approach can find a model set with an optimal size in two steps and ensure a low false alarm rate with acceptable detection rate. It is validated that the approach can be applied for different types of blades, and varied views of each blade type.

The proposed approach outperforms the manual approach, and has been successfully put into practice of the model set selection for turbine blade in the production line, saving a lot of time compared to the manual selection. The approach might be extended for model selection of other reference-based inspection systems.

4 Research Problem II: Adaptive Reference Data Selection for the Assisted Defect Recognition System

4.1 Problem Statement

Parts can vary within the specification during the production process [16, and 53]. For example, for the aluminum die casting, abrasions and wear are common during the lifetime of a mold, and also sand cleaning of the molds leads to variations in wall thickness. These subtle variations are visible in the X-ray images. This leads increasing false alarm rate (increasing rejection rate for defect-free blades), and makes the comparison of older reference image sets with current images under inspection difficult.

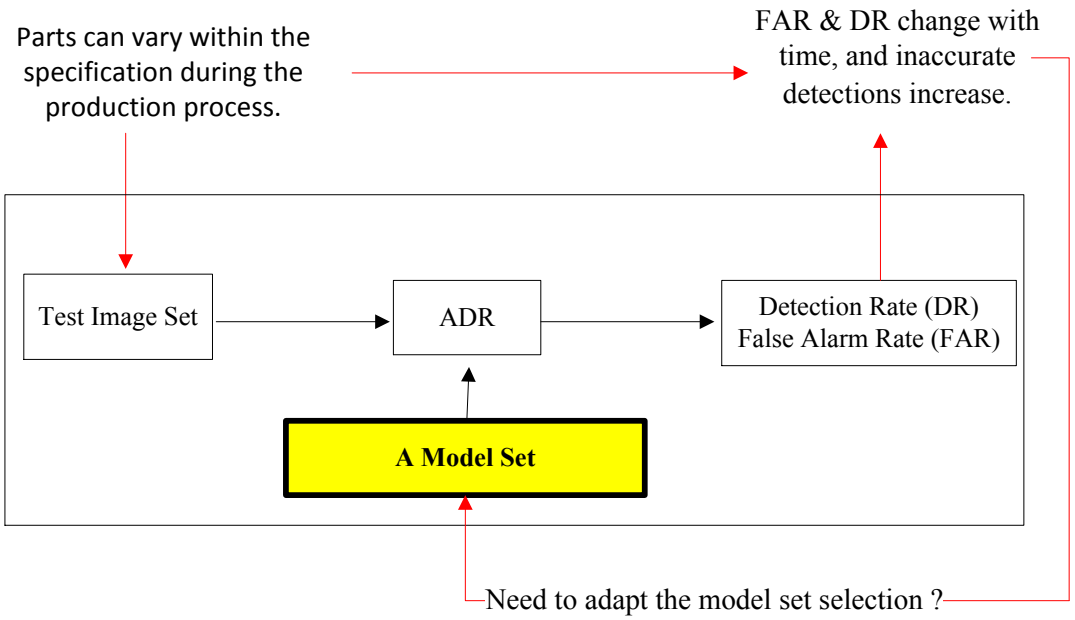


Figure 4.1. Diagram of the problem of reference model set adaptation.

To adapt to parts' variation, the reference model image set should not be static but updated as needed. The diagram of the model set adaption problem is shown in Figure 4.1. The problem is formulated to choose performance metrics to measure and detect the variation, and develop methods of revising the current model

set when significant variation has been detected. The revising method should involve as little human intervention as possible, and find a new model set which can lower the false rejection rate, and maintain an acceptable defect detection rate.

4.2 Proposed Procedure

The framework of the proposed procedure to adaptively select reference model image sets on a timeline is illustrated in Figure 4.2.

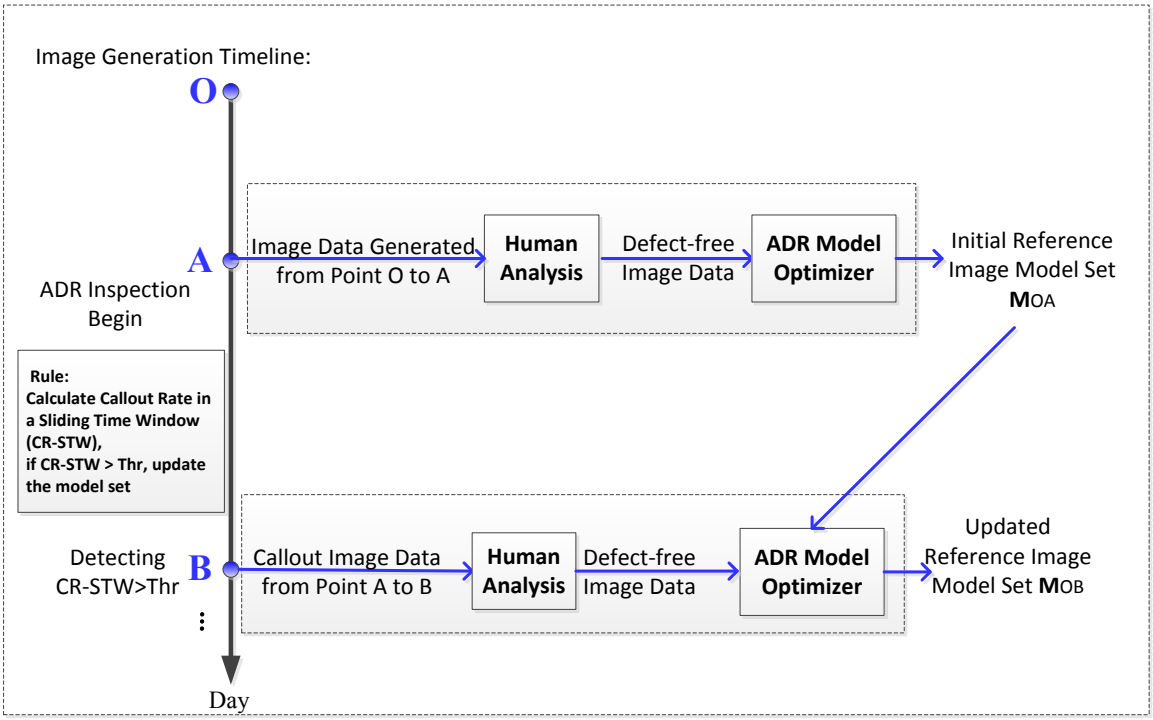


Figure 4.2. Framework of the proposed adaptive method to select the reference model image set.

The procedure first selects an initial reference model set MOA for a time period of OA. Use MOA for ADR to inspect the image data generated after point A, and use the callout rate in a sliding window (CR_STW) to measure the variation of the image data. If callout rate in the sliding window, CR_STW, is greater than a set threshold Thr, CR_STW > Thr, then update the reference model image set MOA to be MOB.

For the selection of MOA, through human analysis of the image data generated from time point O to

point A, obtain the defect-free image data. Feed the defect-free image data into the ADR Model Optimizer, and generate the initial reference model set MOA.

For the selection of the updated reference model set MOB, through human analysis of the callout image data from point A to B, get new defect-free image data. Feed the new defect-free image data augmented with the old reference model set MOA into the ADR Model Optimizer, and generate the new reference model set MOB.

The following will discuss the performance metric of the image data variation CR_STW, and the update method in details.

4.2.1 Callout Rate in a Sliding Time Window

Image data can vary within the specification during the production process. The false alarm rate (false rejection rate) in a sliding time window (FAR_STW) in a time window can be used to measure the variation. However, it is difficult to obtain FAR_STW in practice since FAR_STW equals to the number of false alarm (rejected) images divided by the number of defect-free images, and human experts need to identify the entire truly defect-free image set. To counter this, FAR_STW can be replaced by the callout rate in the sliding time window (CR_STW). FAR_STW is usually approximately proportional to CR_STW since among the callout images, the majorities are false alarm images and the number of truly defective image is very limited.

For CR_STW, the size of the time window should be carefully selected. If it is too small, then CR_STW will only reflect a short-period image data change. If the time window size is too big, then CR_STW cannot reflect the change timely. The size of the time window depends on the specific situation, and can be obtained through extensive experiments.

4.2.2 Model Set Update Method

When significant variation of the image data is detected, the reference model image set needs to be updated. The proposed update method is based on the ADR Model Optimizer. For the ADR Model Optimizer, all

the image data collected need to be analyzed by human experts to obtain the defect-free image data. With the defect-free image data, the ADR Model Optimizer is trained to obtain the reference model set. For the update method, not all the image data but only those callout (rejected) images need to be analyzed by human experts to obtain the falsely rejected images. Those falsely rejected images are defect-free ones, and represent normal variations. With the falsely rejected defect-free image data augmented with the old reference model image set, the ADR Model Optimizer is retrained to generate a new reference model image set.

4.3 Experimental Results and Discussion

This proposed procedure for reference model image sets selection has been validated by X-ray images from Trailing Edge View for turbine blades of blade type "A" through extensive experiments.

A total of 13440 images generated in 122 days are used, including 3 categories: 9835 defect-free images, 235 images with strong indications, and 3370 images with weak indications. The images with strong indications correspond to turbine blade parts that cause safety issues for the jet engines and should be discarded. The images with weak indications correspond to parts with minor anomalies that can be used for the jet engines. Performance metrics used include the callout rate (CR), false alarm rate (FAR), and detection rate (DR) of the images with strong indications and DR of the images with weak indications, respectively, in a sliding-time window. Note that for proprietary information protection the ADR system is tuned arbitrarily, not in the best operating point, and that the results of CR, DR, and FAR of ADR are not actual number in the production line.

Figure 4.3 shows the distribution of the number of images generated each day for each type. From Figure 4.3, we see the number of images generated fluctuates each day for each type. The number of strong defective images is limited compared to the number of the defect-free and weak defective ones.

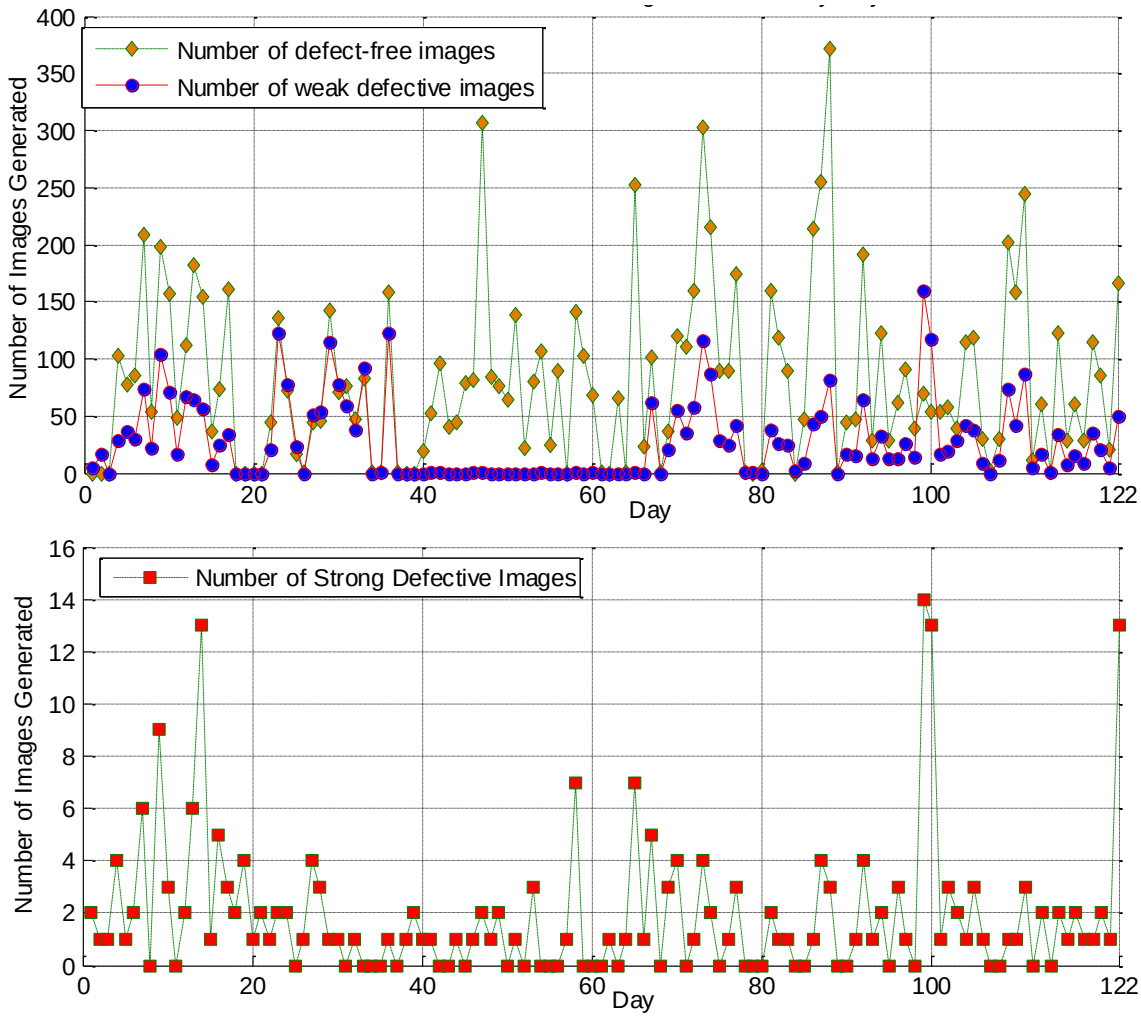


Figure 4.3. Distribution of the number of images generated each day for each type.

The ADR inspection starts with an initial reference model image set MOA. Figure 4.4 shows the performance of the initial reference model image set, with performance metrics including the callout rate (CR), false alarm rate (FAR), detection rate of the strong defective images (DR (SD)) and the weak defective images (DR (WD)) respectively in a sliding time window of 15 days. The initial reference mode image set is selected on Day 37 with about 3000 defect-free images gathered.

Shown as in Figure 4.4, the callout rate observed in a sliding time window of 15 days (CR_STW, the red-filled-circle line) changes after the model MOA has been selected on Day 37. The false alarm rate observed in the same time window (FAR_STW, the blue-up-triangle line) has similar change trend with CR_STW. For DR (SD)_STW (the purple-filled-square line), the value fluctuates, but is acceptable for this

view of the blade type.

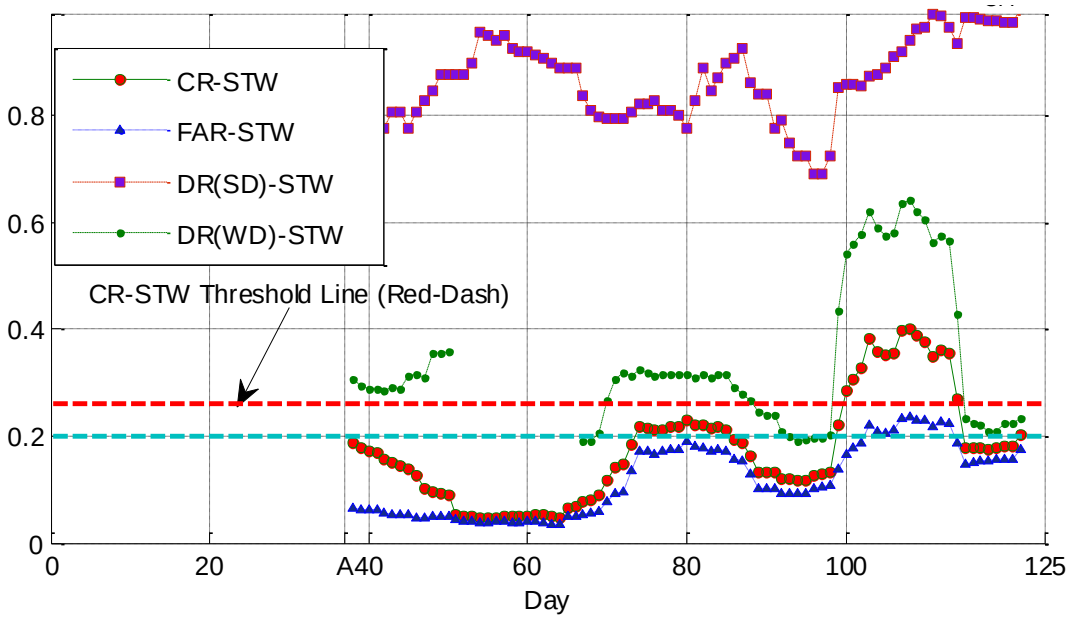


Figure 4.4. Performance of the initial reference model image set MOA.

Note: the CR_STW, FAR_STW, DR (SD), STW and DR (WD)_STW are normalized w.r.t the maximum value of CR_STW.

The DR (WD)_STW is not high, but the WD (weak defective) images correspond to blades with minor anomalies which can be used for jet engines. DR (WD)_STW is not as much concerned as DR(SD)_STW. DR (WD)_STW has no value between Day 51 and Day 66 due to no weak defective images generated during this time interval (see Figure 5). From Figure 4.4, we see CR_STW is above the set threshold (the red-dashed line) on Day 100, and the reference model image set needs to be updated. FAR_STW is in fact above 20% on Day 103.

Figure 4.5 shows the callout rate observed in a sliding time window of 15 days before and after the update of the reference model image set on Day 100 (Point B).

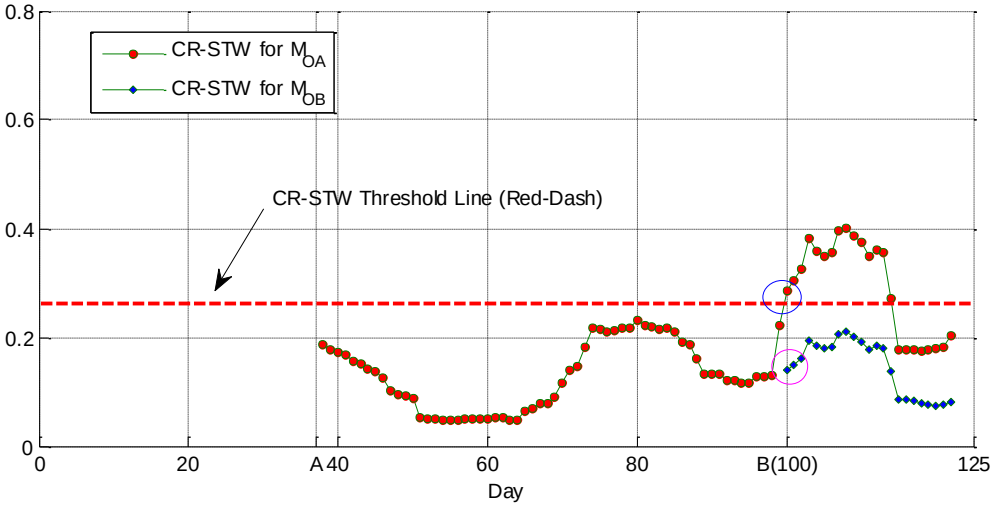


Figure 4.5. Callout rate in a sliding window of 15 days before and after the update of the reference model image set.

Note: CR_STW is normalized w.r.t the maximum DR(SD)_STW for MOA.

From Figure 4.5, we see after updating the model set, CR_STW (the blue-diamond-filled line) decreased below the set threshold (the red-dashed line). The corresponding FAR_STW, DR(SD)_STW are shown in Figure 4.6.

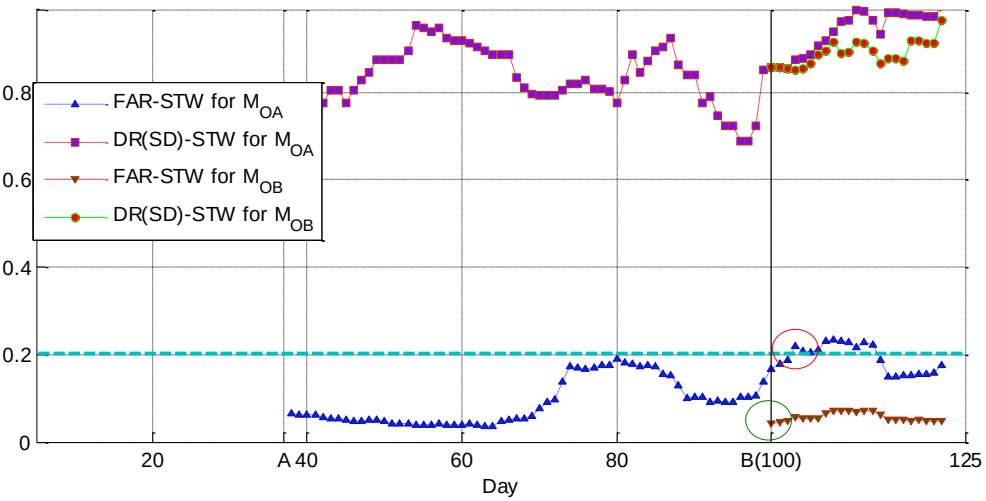


Figure 4.6. False alarm rate and detection rate of strong defective images in a sliding window of 15 days before and after the update of the reference model image set.

Note: FAR_STW and DR(SD)_STW are normalized w.r.t the maximum DR(SD)_STW for MOA.

From Figure 4.6, we see after the model set updated, the FAR_STW is decreased below 10%, far less the than the threshold 20%, and DR(SD)_STW remains in an acceptable level as before the model set update.

For the sliding-time window size, the above experiments use a time window of 15 days. Here we observe the performance metrics of callout rate and false alarm rate in different sizes of sliding time window for the initial reference model set MOA. Figure 4.7 shows the performance of FAR_STW for window sizes of 1, 5, 15, 25 days for MOA.

From Figure 4.7, if the observing window size is too small, like 1-day or 5-day, the fluctuations of FAR are very frequent. If the window size is 25-day, FAR is much stable, but it is not expected that the observing window size too large as it cannot reflect the change timely. To make a trade-off, we select the sliding-time window size to be 15-day.

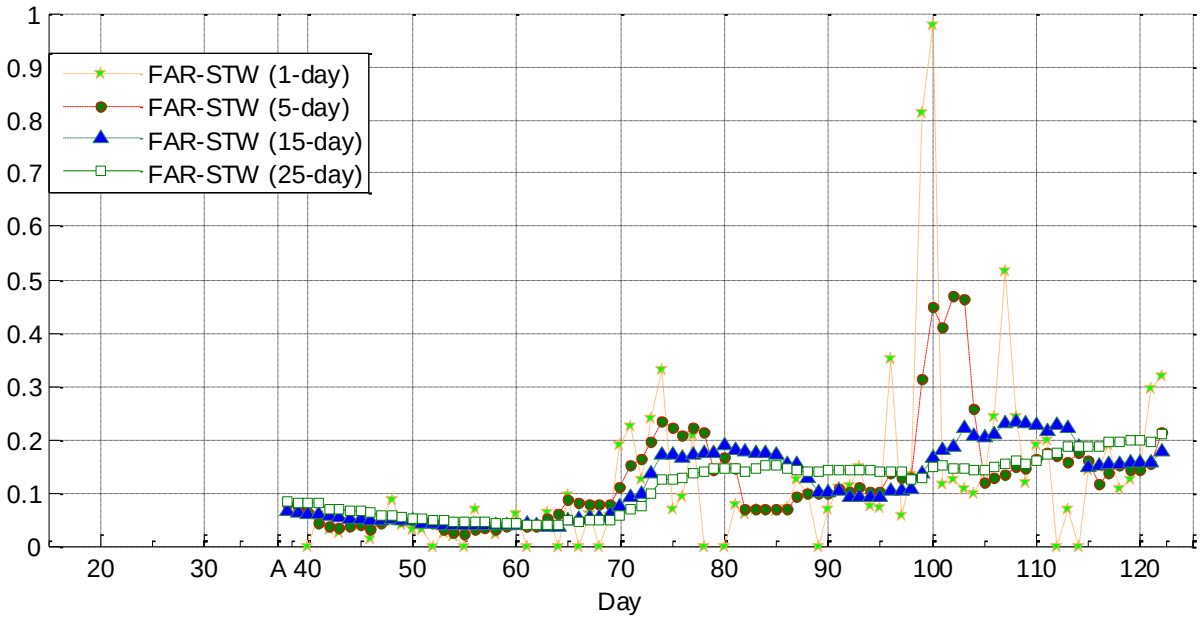
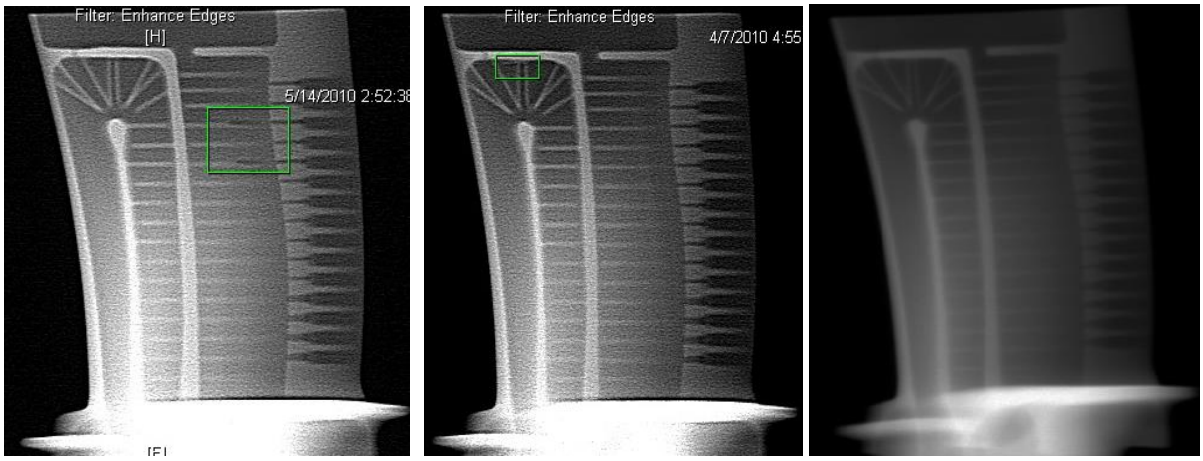


Figure 4.7. FAR in different sliding-time windows for MOA.

Note: FAR_STW are normalized w.r.t the maximum DR(SD)_STW for MOA

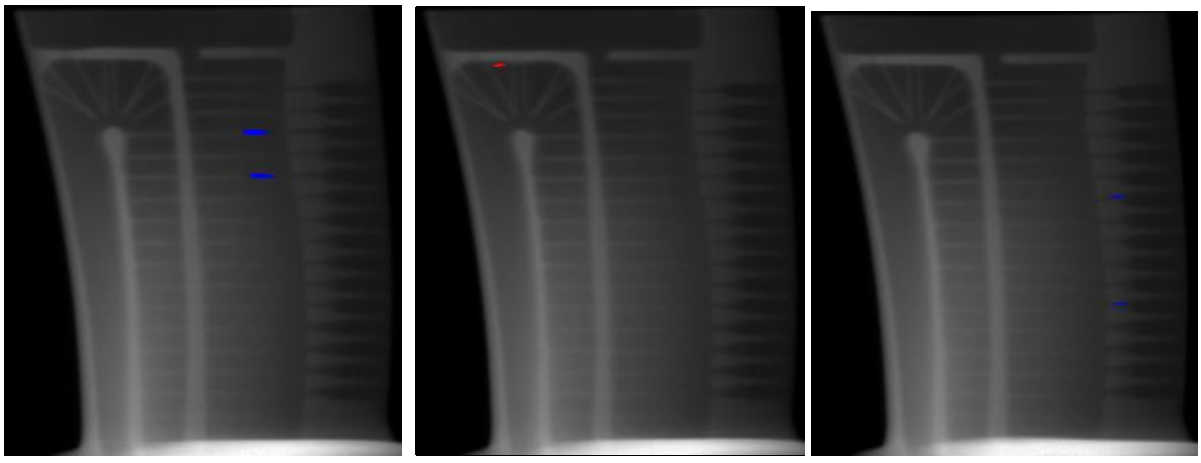
From the above experiments, we see the proposed procedure can adapt the selection of the model set, and the adaption can maintain the ADR performance timely with a low false alarm rate and acceptable detection rate. We further investigate the qualitative results of the detection using ADR based on the model set before and after adaption. Figures 4.8 (a), (b) and (c) show three images with ground truths of defect indications labeled with green bounding boxes by human experts for blade type “B”. Figure 4.8 (a) is an

image with negative defect indications, Figure 4.8 (b) is an image with positive defect indications, and Figure 4.8 (c) is a defect-free image. Figures 4.9 (a), (b) and (c) show the corresponding detection results for images in Figures 4.8 (a), (b) and (c) using ADR based on a selected good model set MOA.



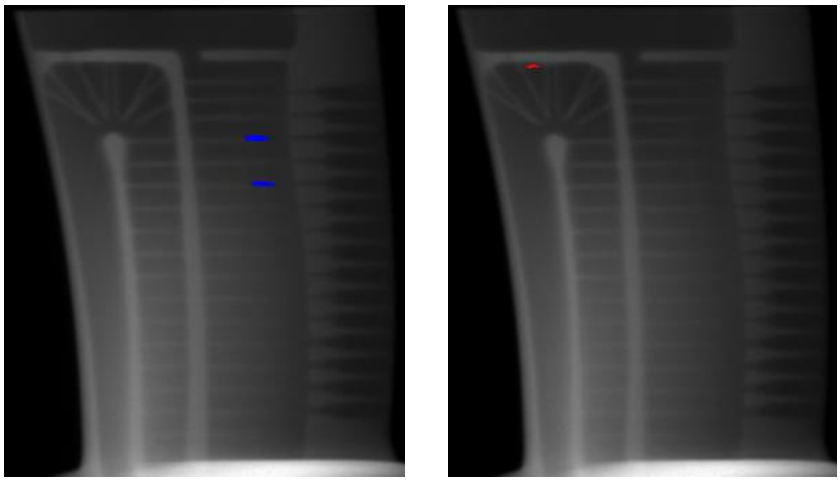
(a) View 1 with negative defects (b) View 1 with postive defects (c) View 1 with no defects

Figure 4.8. Ground truth of defect indications labeled with green bounding boxes in the images by human experts for blade type "B".



(a) View 1 with negative indications (b) View 1 with postive indications (c) View 1 with false indications

Figure 4.9. The corresponding detection results for images in Figure 4.8 using ADR based on the model set MOA.



(a) View 1 with negative indications (b) View 1 with postive indications

Figure 4.10. The corresponding detection results for images in Figure 4.8 using ADR based on the adapted model set MOB.

Note: ADR does not detect any defect in Figure 4.8 (c).

From Figures 4.9 (a) and (b) we see the defect indication type and locations labeled by ADR based on ADR are relatively accurate, and Figure 4.9 (c) is a false alarm. Figure 4.10 shows the corresponding detection results for images in Figure 4.8 using ADR based on the adapted model set MOB. From Figures 4.10 (a) and (b), we see using the adapted model set MOB the defects are still classified and located accurately, and the false alarm image in Figure 4.9 (c) is no longer misclassified.

4.4 Conclusions

We proposed a procedure to adaptively select reference model image sets for the reference based inspection system of turbine blades, ADR. The procedure defines callout rate in a sliding-time window as the performance metric for image data variation. If the callout rate in a sliding-time window is above the set threshold, then the old reference model image set will be updated to adapt to the variation. The update method does not involve much human intervention, and could generate a new reference model image set with much lower false alarm rate than the old set with an acceptable detection rate of strong defective images. From the qualitative results, the updated model set can still detect the truly defects correctly and

remove some false alarms called out by the old model set. The proposed procedure might be extended to reference data selection for other reference-based inspection systems.

5 Research Problem III: Evaluating the Impact of Including Defective Images in the Reference Set on the Assisted Defect System

5.1 Problem Statement

The performance of ADR relies on the reference image data set [15]. The reference data set are selected from a large set of images from good parts. Currently, the selection uses the developed automatic approach – ADR Model Optimizer [15]. However, the precondition to use ADR Model Optimizer is that a large set of good images should be identified by human experts. Due to human's subjectivity, the identified good image set might contain images from parts with minor defects. It is unknown how and to what extent the performance of ADR will be affected if some of these images with defects are included into the reference image set. To deal with this problem, strategies for the following two tasks are needed: 1) including images with defects into the reference data set; 2) evaluating the performance difference before and after the including operation. Figure 5.1 shows the diagram of the problem.

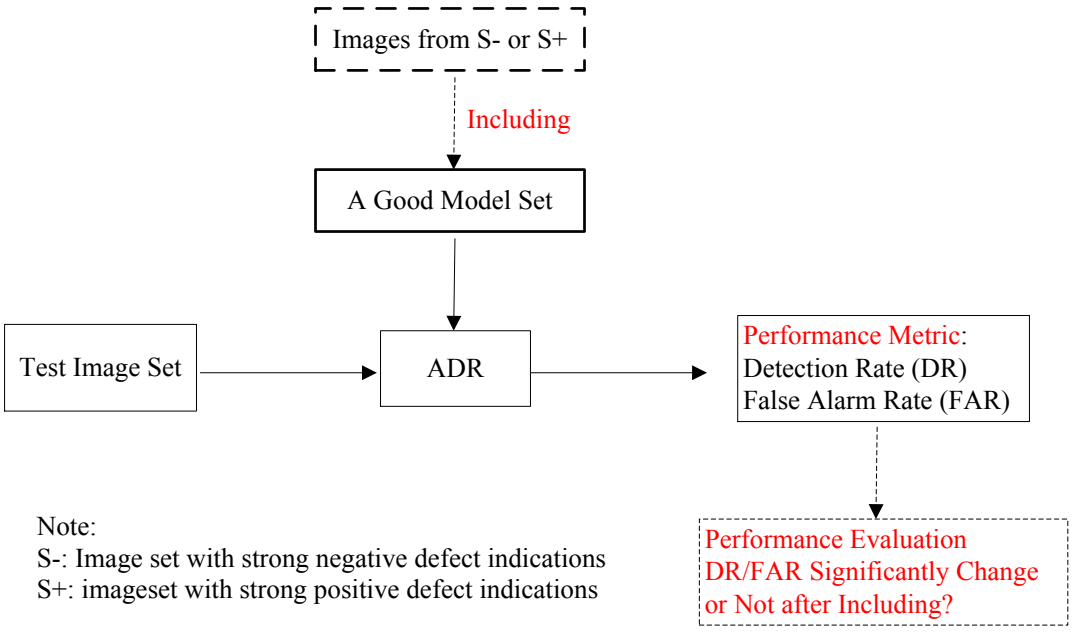


Figure 5.1. Diagram of the problem of including defective images into the model set.

The first task involves how many defective images and which ones to include into the reference image set. For the number of defective images to be included fixed, what strategy to use can generate the worst performance, for example, the lowest detection rate for the images with strong defects. To encounter this task, factors like defect types, locations and defect area might need to be considered in the including process. The rationale is that the reference model set serves as the benchmark of ADR for defect recognition, and it is expected including the worst defective images into the reference model set will degrade the detection rate most greatly. The worst defective images can be defined as the images with the largest defective areas, or with defects of most severe less/excess material, which relates to the defect types, locations and area.

The second task to evaluate the performance change involves statistical comparisons of performance difference before and after the including operation when tested on a given data set. Statistical comparison on a single data set is one of the essential or typical machine learning studies [34]. The appropriate statistical test depends on the setting. For our problem, the detection rates and false alarm rate are the performance metrics to be compared before and after including defective images into the reference image model set. To compare classification rates among multiple models, McNemar's test is the commonly used method [35]. The McNemar's test is a non-parametric statistical test, i.e., it is distribution free. The McNemar's test can be used to determine if there is any significant change for the detection rate or false alarm rate for ADR after changing the reference model set by including defective images.

5.2 Defective Images Inclusion into the Model Set

This section deals with the task of including defective images into the reference image set. Considering factors like defect types, locations and defect area, four methods are proposed to include the defective images into the reference model: single key and multiple key location based, defect area based, and random replacement based methods.

5. 2. 1 Key Locations Identification

Key pixel locations are defined as those pixel locations where common defects occur most frequently based on experts input and ADR indication markings. Based on the defect types, different types of key pixels

locations are identified. Figure 5.2 is an illustration of identifying key pixel locations process for negative images. A good reference model set is selected firstly using ADR Optimizer, and then the strong/weak defective image sets are tested using the selected model set. Based on the indications of ADR callout images with negative defects (173 images), the pixel negative defective histogram is obtained as in Figure 5.2. Select key pixel locations according to the negative defective frequency, indicated as P1, P2, etc., for example, P1 is the most frequently negative defective pixel locations.

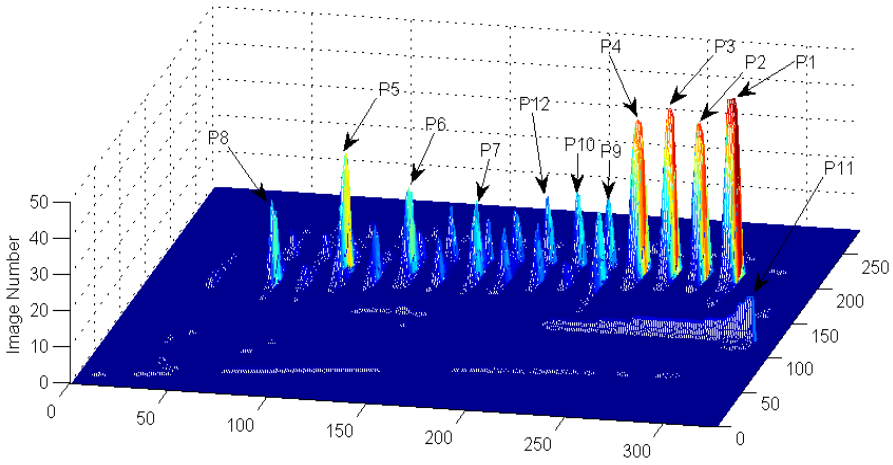


Figure 5.2. Histogram of negative defective pixel locations.

For positive defective images, there are no key pixel locations like the negative defects, which can be observed from Figure 5.3. The positive defect might appear in any location of the blade for 62 positive images under testing. The explanation for this from experts is that the manufacturing process of the blades include a set of hole drillings in fixed locations, which might produce negative defects in these locations.

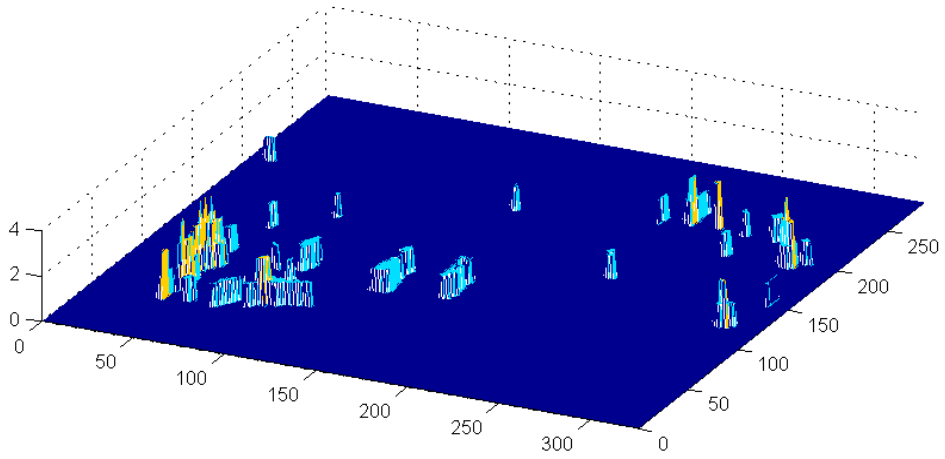


Figure 5.3. Histogram of positive defective pixel locations.

5.2.2 Single/Multiple Key Location Based Inclusion Method

Key locations are those common defective pixel locations for negative-material defect types. For negative defective images, the single key location based inclusion method defines the worst defective images as the ones with the most severe less-material defects at the most frequently occurring negative defective pixel location P1. The method is designed to replace ω best model images with ω worst negative defective images. The best ω model images are the ones which show the least negative-material (highest pixel intensity) at key location P1. The worst ω negative defective images are the ones which show the most extreme negative-material (lowest pixel intensity) at P1.

The multiple key location based inclusion method is similar with the single key location based inclusion method. The difference is that the best ω model images are the ones which show the least negative-material (highest pixel intensity) at key locations P1 through P_ω instead of single pixel location P1. The worst ω negative defective images are the ones which show the most extreme negative-material (lowest pixel intensity) at P1 through P_ω.

5.2.3 Defect Area Based Inclusion Method

The defect area based inclusion method defines the worst negative/positive defective images as the ones with the largest negative/positive defect areas. The method is designed to replace ω best model images with

ω worst negative/positive defective ones. The best ω model images are the ones which show the least total area of defect negative/positive indications. The worst ω negative/positive images are the ones which show the largest total area of defect negative/positive indications.

5.2.4 Random Replacement Based Inclusion Method

To provide a reference for single/multiple key location based and defect area based methods, the random replacement based inclusion method is developed. This method is to replace ω good model images, selected randomly, with ω negative/positive defective images, which are also selected randomly.

The above four methods are designed which defective images to include into the reference image set. For a fixed number of defective images to be included, the aim is to test which method to use that can generate the worst performance. Extensive experiments should be done and results evaluated and compared, for example, the difference of the results of sensitivity (detection rate) before and after including the defective images.

5.3 McNemar's Test

The McNemar's test is used to compare dependent (paired) proportions [34-36, 38]. The test is widely used to compare classification rates (sensitivity, specificity, and overall) among different predictive models. Suppose there are a sample of n defective images for test by ADR using a good reference model image set, Mg, and a non-good set, Mg_d, which includes defective images into Mg. The joint performance of two model sets for McNemar's test can be summarized in a 2x2 table as follows:

Table 5.1. Joint Performance of two model sets 2x2 Table.

		Test Using Reference Model Set Mg_d		
		Success	Failure	Total
Test Using Reference Model Set Mg	Success	A	B	A+B
	Failure	C	D	C+D
	Total	A+C	B+D	n = A+B+C+D

where Mg represents a good reference model set, Mg_d represents the good reference model set Mg with including defective images,

A = Number of defective images which are successfully detected using either of the two model sets, Mg and Mg_d,

B = Number of defective images which are successfully detected using Mg, but failed to detect using Mg_d,

C = Number of defective images which are failed to detect using Mg, but successfully detected using Mg_d,

D = Number of defective images which are failed to detect using both of the two model sets, Mg and Mg_d.

Note that *n* = A+B+C+D, is the number of defective images. It will be assumed that in the underlying test set, π_B and π_C represent the following proportions: $\pi_B = B/(B + C)$ and $\pi_C = C/(B + C)$. If there is no difference between using Mg and Mg_d for detection, the following will be true: $\pi_B = \pi_C = 0.5$. Employing the above information the null and alternative hypotheses for the McNemar's test can be stated [36] as:

$$\text{Null hypothesis } H_0: \quad \pi_B = \pi_C$$

The null hypothesis $\pi_B = \pi_C$ represents the assertion that, given that only one of the reference model sets makes error when used for classification, it is equally likely to be either one. The non-directional alternative hypothesis is

$$\text{Alternative hypothesis } H_A: \quad \pi_B \neq \pi_C$$

The hypothesis is evaluated with a two-tailed test. In order to be supported, the proportion in Cell B, π_B, can be either significantly greater or less than the proportion in Cell C, π_C.

The directional alternative hypothesis is $\pi_B > \pi_C$ or $\pi_B < \pi_C$, which is evaluated with a one-tailed test.

To test H_0, it should only be necessary to examine the test images on which only one of the reference

model sets made an error. No information about the relative performance of Mg and Mg_d is available from the test images on which they are both right and both wrong.

In condition on the number of test images K = B+C where only one reference model set made an error, then for the observed K = k, C follows a binomial distribution B(k,1/2) distribution under H_0 [38]. The null hypothesis is thus tested by applying to a two-tailed distribution to the observation of a random variable M drawn from a B(k,1/2) distribution:

$$P=\begin{cases} 2\Pr(C\le M\le k), when\ \ C>k/2 \\ 2\Pr(0\le M\le C), when\ \ C<k/2 \\ 1.0, when\ \ C=k/2 \end{cases} \tag{5.1}$$

The probabilities can be computed directly as follows:

$$P=\begin{cases} 2\sum_{m=C}^{k}\binom{k}{m}\left(\frac{1}{2}\right)^{k}, when\ \ C>k/2 \\ 2\sum_{m=0}^{C}\binom{k}{m}\left(\frac{1}{2}\right)^{k}, when\ \ C<k/2 \\ 1.0, when\ \ C=k/2 \end{cases} \tag{5.2}$$

or alternatively, tables of the Binomial distribution may be used. H0 is rejected when P is less than some chosen significance level α/2. If k is large enough (k > 50) and C is not too close to k or 0, a normal approximation to the exact Binomial probability can be used [36]. Under H_0 and conditional on K =k, the expectation of C, E(C) = 1/2, and the variance of C, Var(C) = k/4. Then define the statistic

$$W=\frac{\left|C-k/2\right|-1/2}{\sqrt{k/4}} \tag{5.3}$$

which should be approximately N(0,1) under H_0. Compute the P-value P = 2Pr (Z≥ w) (where Z is a random variable with distribution N(0,1) and w is a realized value of W), and reject H0 if P< α /2, where α is a chosen significance level. The -1/2 in the numerator of equation 5.3 is a continuity correction factor [39]. This latter form of the test is equivalent to the χ^2 test of McNemar [40].

5.4 Experimental Results and Discussion

This section presents the experiments of using the above four methods to include negative and positive defective images into the good reference model set for ADR. The results of the detection rates and false alarm rate are compared with using the good reference model set, and McNemar's test is used to determine if there is any significant change and how many defective images can be tolerated in a reference model set without changing the performance of ADR. Note that for proprietary information protection the ADR system is tuned arbitrarily, not in the best operating point, and that the results of DR and FAR of ADR are not actual number in the production line.

The test image data used is from View 1 of the blade type X, including 9835 good images and 235 strong defective images (images with strong defect in the corresponding turbine blades). The 235 defective images consist of 173 negative defective images and 62 positive defective images. A good reference image set, Mg, consisting of 130 images is selected from the 9835 good images using the ADR Model Optimizer. The detection rates (DR) and false alarm rate (FAR) using the good reference model set are listed in Table 5.2.

Table 5.2. ADR performance using the good reference model set Mg.

Performance Metrics	DR	FAR
Using Mg	84.68%	4.50%

From Table 5.2, we see the detection rate is 84.68%, while the false alarm rate is only 4.50%. Negative defective images are included into the good set Mg using the four proposed methods in Section 5.2, and the image data set are tested using the reference model sets with defective images. Table 5.3 lists the results for the reference model sets with including 5 negative defective images.

Table 5.3. ADR performance using the reference model sets with 5 negative defective images.

Inclusion Methods	DR	FAR
Single Location Based	77.45%	4.53%
Multiple Location Based	80.43%	4.49%
Defect Area Based	82.13%	4.46%
Random Replacement Based	81.62%	4.54%

Note: the results of using random replacement based method are obtained by averaging 5 runs.

From Table 5.3, we see that the FAR for each of the four methods is less than 5%, which is acceptable in practice. For detection rate, the location based methods have worse performance than the defect area and random replacement based methods, especially the single location based method. Table 5.4 lists the results for the reference model sets with including different number of negative defective images using the Single Location Based Inclusion method.

Table 5.4. ADR performance of including negative defective images using singe location based method.

No. of negative defective images included	DR	FAR
1	83.40%	4.45%
2	83.83%	4.43%
3	82.55%	4.54%
4	82.13%	4.58%
5	77.45%	4.53%
10	74.89%	4.69%

From Table 5.4, we see as the number of negative defective images included into the reference model set increases, the detection rate decreases, but FAR has no change trend. Including negative defective images will impact the detection rate. To evaluate if the impact on detection rate is significant or not, McNemar's

test can be used. Table 5.5 is the joint performance of the good reference model set Mg and the reference model set with 10 negative defective images included using single location based method.

Table 5.5. Joint performance of the good reference model set Mg and the reference model set with 10 negative defective images included using single location based method 2x2 Table.

Testing defective image set		Test Using Reference Model Set Mg_d with 10 negative defective images		
		Success	Failure	Total
Test Using Reference Model Set Mg	Success	175	24	199
	Failure	1	35	36
	Total	176	59	235

From Table 5.5, we have C = 1, k = 25, and using equation 5.2, we obtain P = 1.5e-6. If set the significance level $\alpha = 0.05$, then $P << \alpha/2$, indicating that the detection rate is significant different after including 10 worst negative defective images at location P1 into the good reference model set Mg. In fact, the detection rate decreases from 199/235 = 84.68% to 176/235 = 74.89%, nearly a 9.8% drop as shown in Table 5.4. Similarly, McNemar's test is performed for model sets with 1, 3, and 5 worst negative defective images at location P1. Table 5.6 lists the results of McNemar's test.

Table 5.6. McNemar's test for model sets with 1, 2, 3, 4, 5, and 10 worst negative defective images at the single location P1.

No. of negative defective images included	DR	P value of McNemar's test	Significant Difference (with $\alpha/2 = 0.025$)
1	83.40%	0.25	No
2	83.83%	0.50	No
3	82.55%	0.063	No
4	82.13%	0.031	No
5	77.45%	1.5e-5	Yes
10	74.89%	1.5e-6	Yes

From Table 5.6, including 3 or less negative defective images, the detection rate does not change significantly. With 5 or more negative defective images in the model set, the detection rate significantly decreases.

The following focuses on the inclusion of positive defective images. Positive defective images are included into the good model set using the random replacement based method. Table 5.7 lists the results for the reference model sets with including different number of positive defective images.

Table 5.7. ADR performance of including positive defective images.

No. of positive defective images included	DR	FAR
1	83.40%	4.45%
3	82.98%	4.48%
5	82.55%	4.35%
10	82.98%	4.82%
20	82.13%	4.97%
25	84.25%	5.47%

From Table 5.7, we see that the FARs for model sets with 1 to 20 positive defective images are less than 5%, which is acceptable in practice. The detection rate does not have any obvious change, and McNemar's test is applied to evaluate the change in the following. Table 5.8 lists the results of McNemar's test.

Table 5.8. McNemar's test for model sets with 1, 3, 5, 10, and 20 positive defective images.

No. of positive defective images included	DR	P value of McNemar's test	Significant Difference (with $\alpha/2 = 0.025$)
1	83.40%	0.25	No
3	82.98%	0.13	No
5	82.55%	0.13	No
10	82.98%	0.063	No
20	82.13%	0.34	No

From Table 5.8, including up to 20 positive defective images, there is no significant change for the detection rate. The inclusion of positive defective images into the good model set has no significant impact for the performance of ADR.

To further investigate why including certain number of negative defective images into the good model set would affect the total DR, the detection rate for positive and negative images are calculated separately. Table 5.9 shows the DRs for positive images and negative images with and without including negative defective images using single location based method.

Table 5.9 ADR performance of including negative defective images.

No. of negative defective images included	DR(-)	Significant Difference (with $\alpha/2 = 0.025$)	DR(+)	Significant Difference (with $\alpha/2 = 0.025$)
0	167/173 = 96.53%	No	30/62 = 48.39%	No
1	166/173 = 95.95%	No	30/62 = 48.39%	No
3	165/173 = 95.38%	No	29/62 = 46.77%	No
5	153/173 = 88.44%	Yes	29/62 = 46.77%	No
10	146/173 = 84.39%	Yes	30/62 = 48.39%	No

Note: DR(-) represents detection rate for negative defective images and DR(+) for positive defective images

Seen from Table 5.9, with increasing the negative defective images in the model set, the DR for negative defective images decreases rapidly and becomes significant, but the DR for positive defective images is nearly not affected. This is reasonable because including the negative images with the worst negative defective pixel indication at P1(279, 207) into the model set would let ADR learn the statistics of these negative defective images, especially at the key location P1, which is the location with the most frequent negative defect occurring, and thus classify images with similar negative defect indications as non-defective, and decrease the overall detection rate. This is evident from the difference between the histograms of negative defective pixel markings before and after including the negative defective images. Figure 5.4 shows the histogram of negative defective pixel markings on the image by ADR using the good model set,

and Figure 5.5 shows the histogram of negative defective pixel markings on the image after including five negative defective images into the good model set using single key location based method, and Figure 5.6 is the difference of the histograms before and after the inclusion.

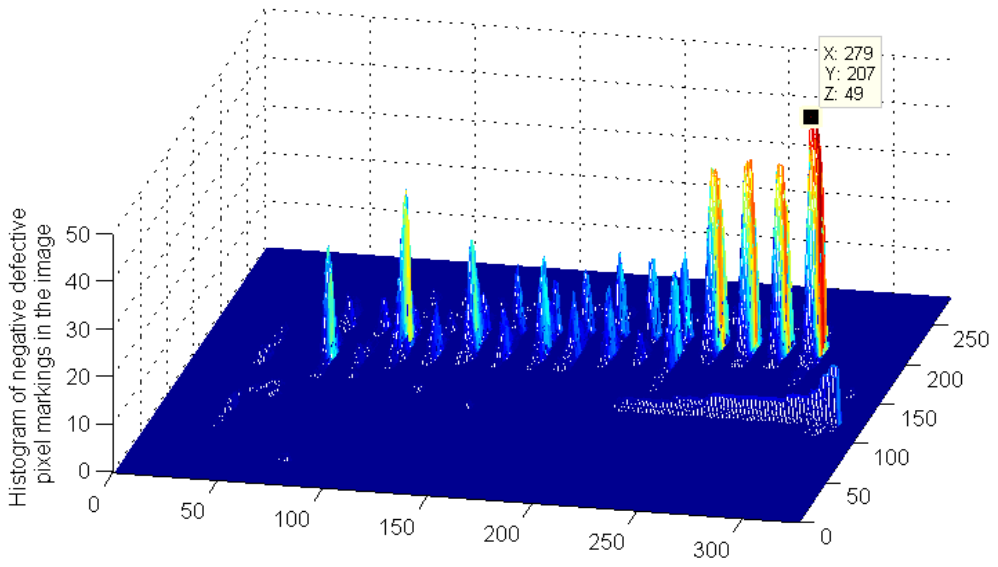


Figure 5.4. Histogram of negative defective pixel markings in the image using the good model set.

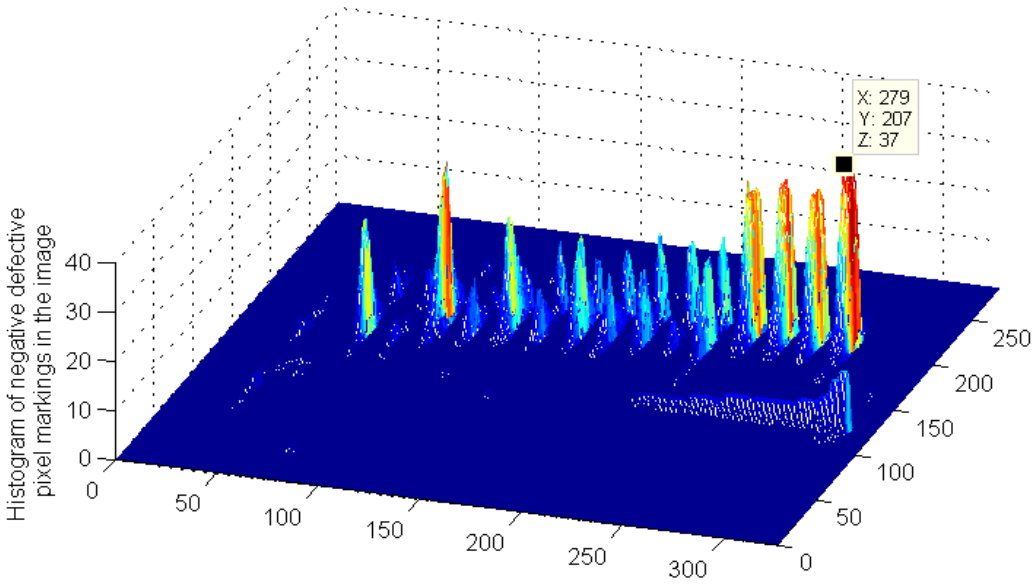


Figure 5.5. Histogram of negative defective pixel markings in the image after the including 5 negative defective images into the model set.

We see, before the inclusion the number of negative defective images with negative defective marking at P1 by ADR is 49, after the inclusion this number drops to 37, with 12 out of the 49 images considered non-defective and not marked at P1 as shown in Figure 5.6. This approximately matches the results of Table 5.9. From Table 5.9, we see the detection of the negative defective images before and after including five negative defective images are 167 and 153 (highlighted by the yellow color), with a total of 14 images being considered non-defective after the inclusion. Out of the 14 images, 2 images are not labeled negative defective at P1. This might be due to the fact that the included defective images have defect indications at other pixel locations besides P1. Figure 5.7 (a), (b) and (c) show three of the fourteen images with defect indications labeled with green bounding boxes by human experts.

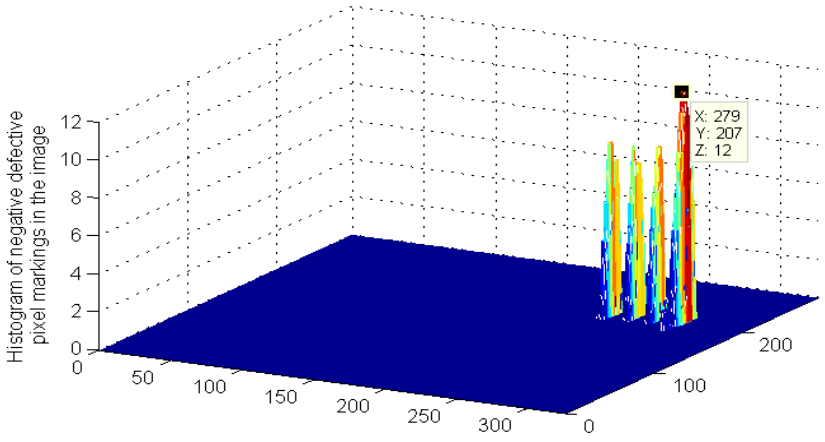
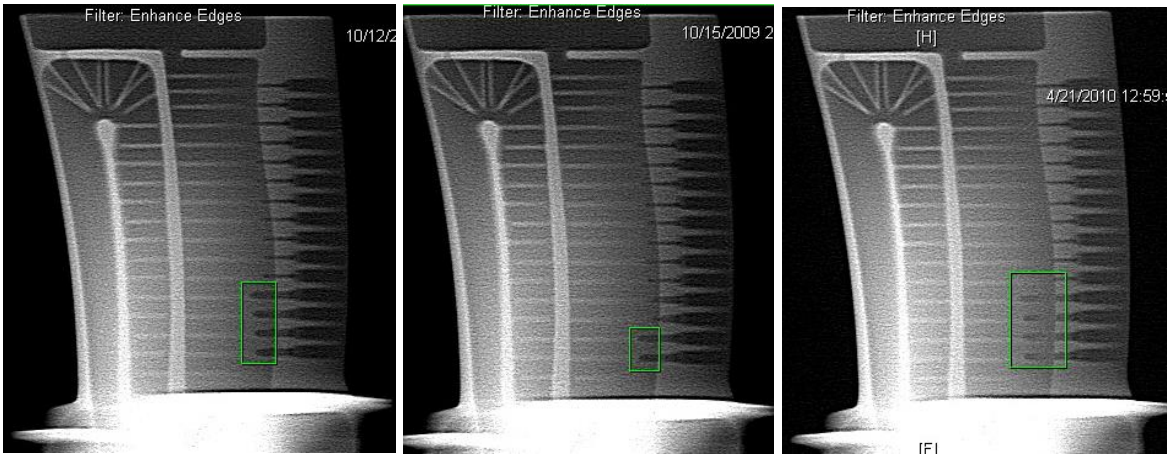


Figure 5.6. Difference of histograms of negative defective pixel markings before and after the including 5 negative defective images into the model set.

Figure 5.8 (a), (b) and (c) show the corresponding detection results for images in Figure 5.7 (a), (b) and (c) using ADR based on the good model set, but not detected by the model set with five negative defective images included. In Figure 5.8 (a) and (b), P1 is marked with negative defect by the good model set, matched the ground truth in Figure 5.7 (a) and (b). In Figure 5.8 (c), P1 is not marked, but the pixels close to P1 are marked, and the markings are within the bounding box of the ground truth in Figure 5.7 (c). After including the negative defective images into the model set, the defects of images in Figure 5.7 and 5.8 are

not detected by ADR, and thus we can see the destructiveness aspect of the inclusion of negative defective images.



(a) (b) (c)

Figure 5.7. Ground truth of strong negative defect indications labeled with green bounding boxes in the images by human experts for blade type "B".

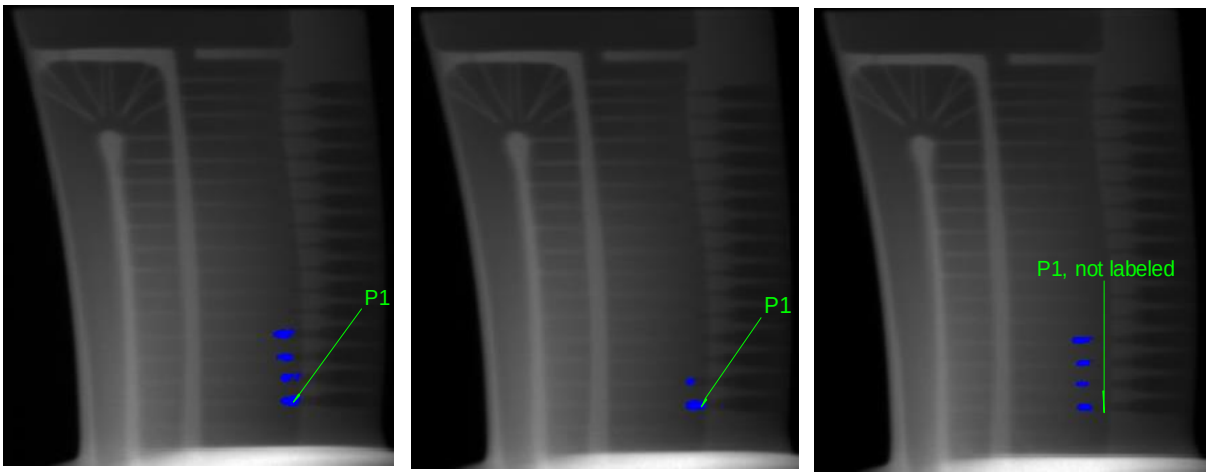


Figure 5.8. The corresponding detection results for images in Figure 5.7 using ADR based on the good model set, but not detected by the model set with including negative defective images.

Table 5.10 shows the DRs for positive images and negative images with and without including positive defective images. Seen from Table 5.10, as the number of included positive defective images increases up to 10, the detection rate for negative images is almost the same and not significantly affected. The change

of detection rate for positive defective images is slight and not significant. A possible reason could be that the positive defect indications could occur everywhere in the turbine blade, and there is no frequent defective pixel locations, and thus including these images into the good model set might not be as damaging as negative defective images.

Table 5.10. ADR performance of including positive defective images.

No. of positive defective images included	DR(-)	Significant Difference (with $\alpha/2 = 0.025$)	DR(+)	Significant Difference (with $\alpha/2 = 0.025$)
0	167/173 = 96.53%	No	30/62 = 48.39%	No
1	167/173 = 96.53%	No	30/62 = 48.39%	No
3	166/173 = 95.95%	No	29/62 = 46.77%	No
5	166/173 = 95.95%	No	28/62 = 45.16%	No
10	166/173 = 95.38%	No	30/62 = 48.39%	No

5.5 Conclusions

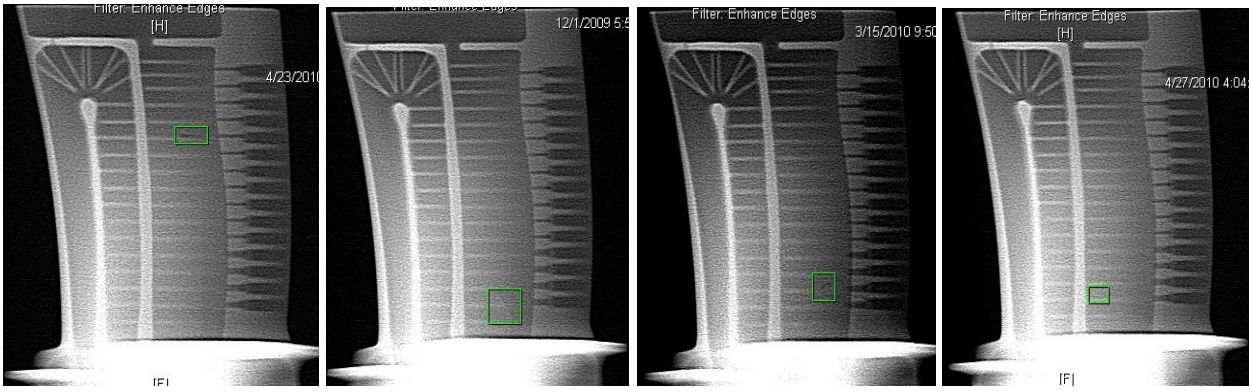
The paper studies the impact on the performance of a reference based industrial part inspection system of including defective images into the reference image data Four methods to include the defective images are proposed based on the defect types, defect area and locations. McNemar's test is used to evaluate if the impact is significant or not. Experimental results show that including negative defective images based on key defect locations can generate the worst performance as to the detection rate. When the number of negative defective images increases, the detection rate decreases. The decrease is significant when the number is 5 or more based on the McNemar's test. Including positive images up to 20 has no significant impact on the detection rate. Further investigation shows that including negative defective images mainly affects the detection rate of the negative images, but not the detection of positive defective images. Including positive defective images has no significant impact on the detection of positive defective images or negative defective images. The possible explanation could be that the negative defect might occur

frequently in some key locations, but the case for positive defect indication does not hold, and thus including positive defective images into the good model set might not be as destructive as the negative defective images. More investigation is needed to further verify this conclusion. Besides, the results also show that the false alarm rate is acceptable for defective images of any type included. The experimental results provide guidance for practice that the reference image data could tolerate limited number of defective images for reference based inspection systems.

6 Research Problem IV: Model-based Approach to Automated Defect Recognition Using Simple Features

6.1 Problem Statement

Turbine blades are one of the basic components of a gas turbine, and are responsible for extracting energy from the high temperature, high pressure gas produced by the combustor [54]. The quality specifications of turbine blades are very demanding, and imperfections of the internal structure, including inclusions and holes may put blade lifetime at risk, and thus make turbine blades the most sophisticated industrial parts for inspection. Defect indication of the blades can be divided into four types: strong negative, weak negative, strong positive and weak positive defect indications. Figure 6.1 shows sample images with the four types of defects labeled with green bounding boxes for blade type "B" from View 1.



(a) Strong negative (b) Strong positive (c) Weak negative (d) Weak positive

Figure 6.1. Sample images with four types of defect labeled with green bounding boxes or blade type "B" from View 1.

The images with strong defect indications corresponds to bad blades which cannot be used in making a turbine and must be deposed in the production, especially, the negative defects which are the most damaging. The images with weak defect indications corresponds to the blades that have minor anomalies and can be tolerated in production.

ADR is used to detect these defects. However, like any system, ADR is not ideal and there are still missed detections, for example, for the View 1 of blade type "B", the detection rate is 167/173 = 96.53%. The objective is to develop a method to find the undetected images by ADR without increasing any false rejection rate.

6.2 Proposed Approach

ADR creates individual statistical models at each pixel based on the reference model image set. However, defects can be viewed as irregular objects and occur as groups of pixels. Negative defects tend to appear in regions where holes are drilled. And turbine blades are produced in high volume, and methods to improve the ADR for detection should be fast in implementation. Based on these observations and facts, we propose an approach for defect recognition using simple features of scan line and modified Haar-like features. The Diagram of the proposed approach is shown is Figure 6.2.

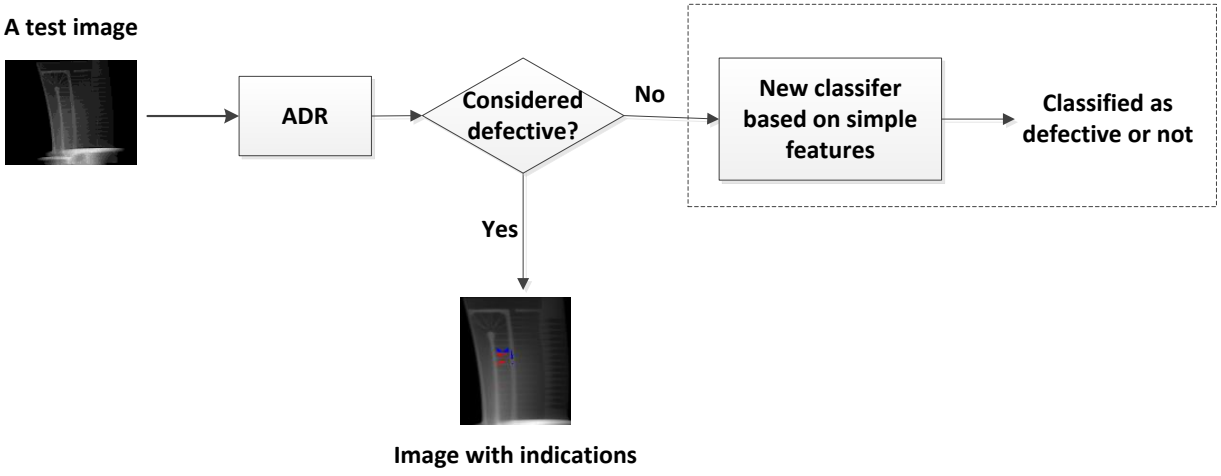


Figure 6.2. Diagram of the proposed approach.

As shown in Figure 6.2, the proposed approach combines the ADR and a new classifier. A test image is first fed into ADR. If considered defective, the test image would be called out with indications by ADR. Otherwise, the image would be then fed into the new classifier to be classified as defective or not.

The diagram of the new classifier is shown in Figure 6.3.

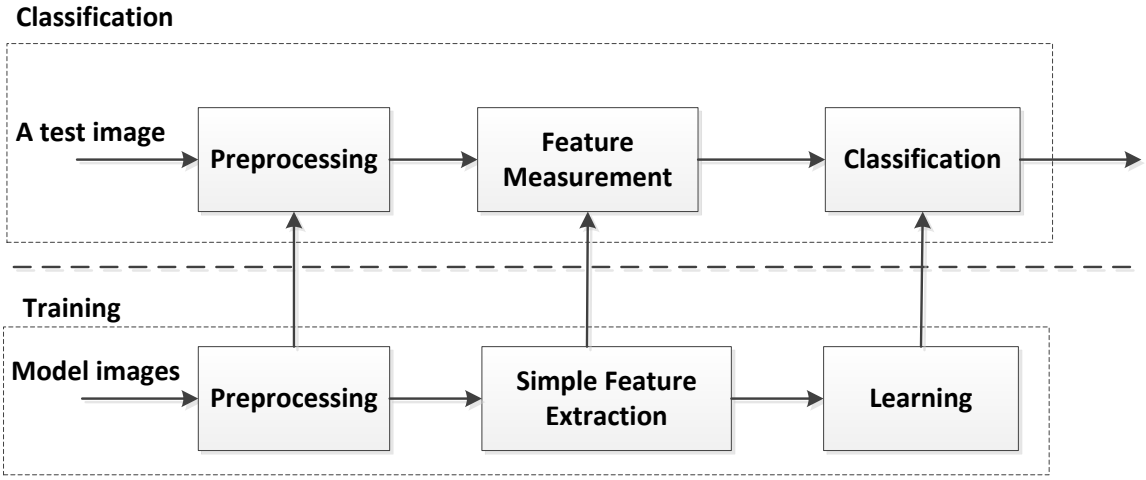


Figure 6.3. Diagram of the new classifier based on simple features.

As shown from Figure 6.3, like most statistical pattern recognition systems [41], the new classifier is operated in two modes: training (learning) and classification (testing). The preprocessing module consists of image registration and normalization, which we adopt the similar methods used in the ADR system. In the training mode, scan line and modified Haar-like features are extracted from the reference model images and a normal feature space is defined. In the classification mode, the classifier determines if the test image is defective or not based on the measured features.

6.3 Image Preprocessing

Image preprocessing techniques including registration and normalization are used to factor out those changes of image brightness which are irrelevant to defects [4], such as blade misalignment and attenuation differences due to source and detector gain variations during imaging.

6.3.1 Image Registration

Image registration is the process of transforming different image data into one coordinate system. Data may be from different depths, viewpoints, positions, etc. [63]. The goal of image registration is to rectify the minor pose variations in the blade positioning.

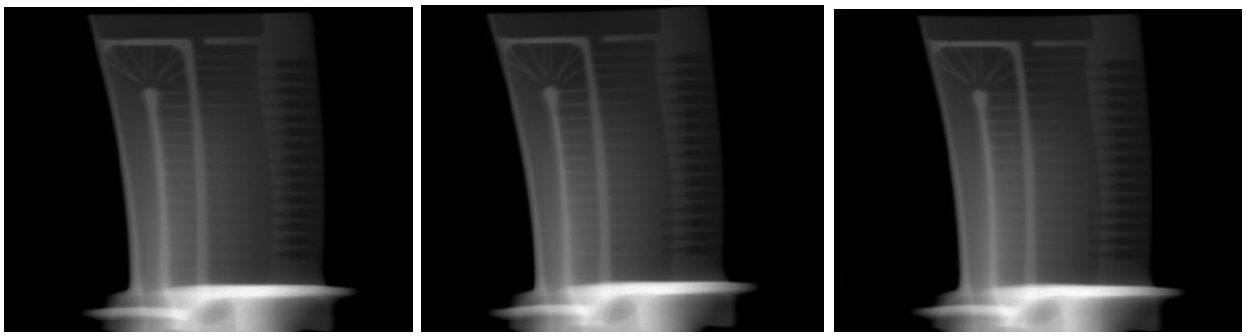
Given a test image B and a template image A, registering B to A can be formulated as:

$$T^* = argmin_T \mathrm{D}(\mathrm{A}(\mathrm{i},\mathrm{j}), \mathrm{B}(\mathrm{T}(\mathrm{i},\mathrm{j}))) \quad (6.1)$$

The aim is to find the optimal geometric transform T^* that transform B into the spatial alignment with A, where (i, j) is a pixel location on the images [4]. The geometric transform includes translation, rotation, etc. $\mathrm{D}(\mathrm{A}(\mathrm{i},\mathrm{j}), \mathrm{B}(\mathrm{T}(\mathrm{i},\mathrm{j})))$ is the image similarity measure between A and the transformed B. The widely used similarity measure is mutual information (MI). MI is not sensitive to illumination and appearance changes, to counter this, disjoint information is used as the similarity measure for registration [49]. Disjoint information is defined as the joint entropy excluding the mutual information

$$D(A,B) = H(A,B) - I(A,B) \quad (6.2)$$

where $\mathrm{D}(\mathrm{A},\mathrm{B})$, $\mathrm{H}(\mathrm{A},\mathrm{B})$, and $\mathrm{I}(\mathrm{A},\mathrm{B})$ represent the disjoint information, the joint entropy and mutual information between A and B respectively. Figure 6.4 gives an example of image registration.



(a) The template image A (b) The image to be registered B (c) The image after registration T*(B)

Figure 6.4. Image registration example.

From Figure 6.4, we see the image B is slightly inclined to the left compared to the template image A, and after registration it is upright.

6.3.2 Image Normalization

Minor part variations, positioning and attenuation differences due to source and detector gain variations might cause image spatial appearance shift. Image normalization is to compensate such variations [64]. Images are registered to the template image as before normalization.

The image normalization [4] uses the following operation:

$$I'\big(T(i,j)\big) = \mathrm{I}\big(T(i,j)\big) - F(\mathrm{I}\big(T(i,j)\big)) + I_0(i,j) \tag{6.3}$$

where I_0 is the baseline image, and is generated using 1D median filter at pixel (i, j) for the model images after a large radius 2D median filter F is applied to these images, and I(T(i,j)) represents the pixel (i, j) in the registered image.

A test image I is first spatially transformed to a template images as I(T) in the registration process. The difference between I(T(i,j)) and its spatially median filtered version F(I(T(i,j))) at pixel (i, j) is then used to compensate $I_0(i,j)$ in the baseline image.

6.4 Feature Extraction

An ideal feature extractor should produce representations of defects to be distinctive from the normal structure [42]. The feature extraction is required to be computationally cheap in order to be rapid to cope with the industrial inspection [43]. ADR uses the low-level feature – the pixel intensity to build a PDF at each pixel location by Parzen window density approximation based on the model image set. Due to the part-to-part variation and image-to image variation, for example, image appearance change due variations between the imaging source and detector. To use intensity at each pixel alone to create statistical models by ADR is not sufficient to discriminate the defects.

We investigate the defective images of turbine blades labeled by human experts. Figure 6.5 shows four strong negative defective images with the defects labeled with green bounding boxes. We see defects are irregular in shape, and tend to appear as groups of pixels.

The defective pixels are different in intensity from their neighboring pixels. Based on these observations and inspired by [44], where a scan line is used to measure the statistics of the intensity along the line, a modified scan line feature is proposed here for defect recognition in some particular region. Inspired by [45-46], which uses Haar-like features for rapid object detection, modified Haar-like features are proposed.

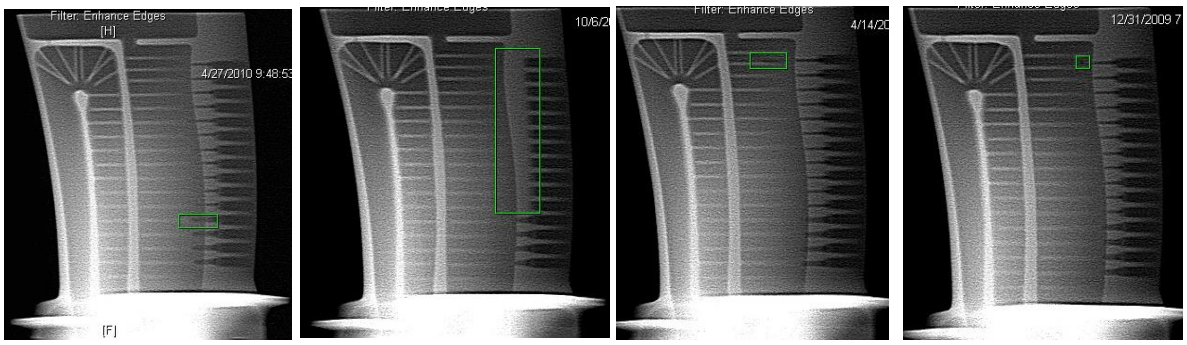


Figure 6.5. Four strong negative defective images with the defects labeled with green bounding box.

6.4.1 Scan-line Feature

The scan line gray-level profile feature is used for micro-crack defect detection by [44]. Inspired by this, we use it for blade defect detection. As shown in Figure 6.6(a), the scan line is a vertical line (red color) across the region of interest (labeled by the green bounding box), Figure 6.6 (b) shows the gray-level profiles of the scan line for three images, one is defective in the green bounding box, and the other two are defect-free. From Figure 6.6 (b), we see the difference between the defective image and the other two defect-free images based on the corresponding gray-level profiles of the scan line.

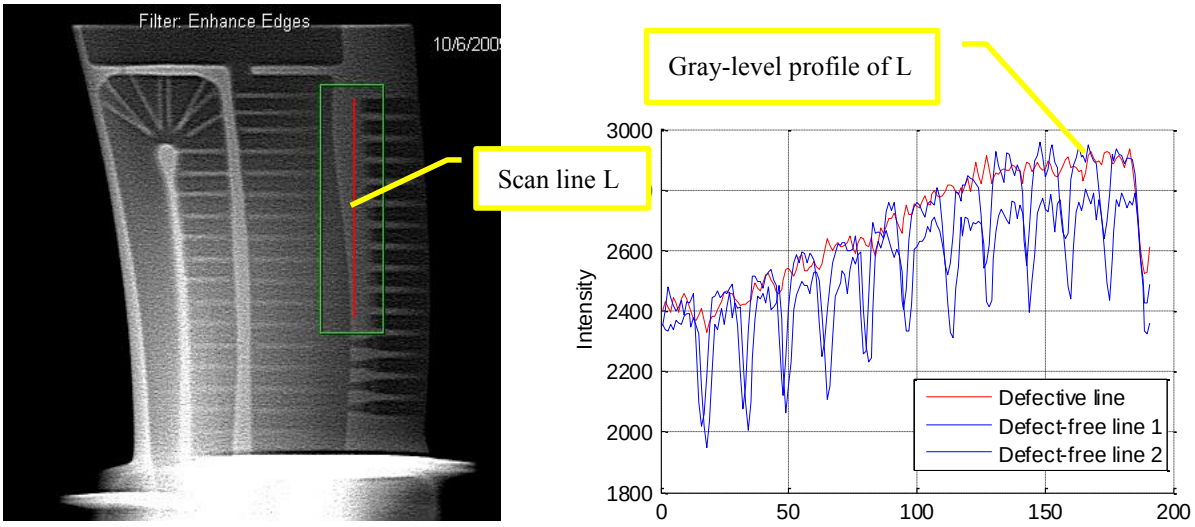
Let L represent gray-level profile of the scan line, and L' represent normalized gray-level values on the scan line. The two feature measures average and standard deviation of L' are employed:

$$\mathrm{f}_1\colon\ \mu = \frac{1}{N}\sum_i^N L_i' \tag{6.4}$$

$$\mathrm{f}_2\colon\ \ \sigma = \left(\frac{1}{N-1}\sum_i^N (L_i' - \mu)^2\right)^{1/2} \tag{6.5}$$

where,

$$L_i' = \frac{L_i - L_{min}}{L_{max} - L_{min}} \tag{6.6}$$



(a) A image with defects in the green box (b) Gray-level profiles of the scan line for three images

Figure 6.6. Scan line feature extraction.

$$L_{min} = \min(L_i), for\ i = 1, \dots, N \tag{6.7}$$

$$L_{max} = \max(L_i), for\ i = 1, \dots, N \tag{6.8}$$

where L_i refers to the intensity of the i^{th} pixel of the scan line L, and N is the total pixel number on L.

μ is a measure of central tendency for roughly symmetric distribution of the normalized scan line signal. σ is a measure of statistical dispersion.

6.4.2 Haar-like Feature

Haar-like features were proposed by Viola and Jones to be used in the real-time face detector [45]. There are two motivations for using Haar-like features: one is that it can act to encode ad-hoc domain knowledge which is not easy to learn from limited training data, and the other is fast in computing using integral images [46].

Viola and Jones used three simple Haar-like features: two-rectangle feature, three-rectangle feature, and four-rectangle features [46], as shown in Figure 6.7. The two-rectangle features can represent the edge features, three-rectangle the line features and four-rectangle special diagonal line features. Lienhart [65] extended the Haar-like features by an efficient set of 45 degree rotated features, which can use additional domain-knowledge that is hard to learn to the learning framework. The extended features are shown in Figure 6.8 with four edge features, eight line features and two center surround features. Using the extended features, false alarm rate for face detection is decreased significantly.

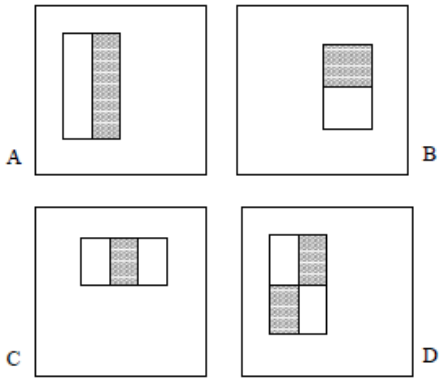


Figure 6.7. Original Haar-like features. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. (A) and (B) are two-rectangle features, (C) and (D) are three- and four-rectangle features respectively.

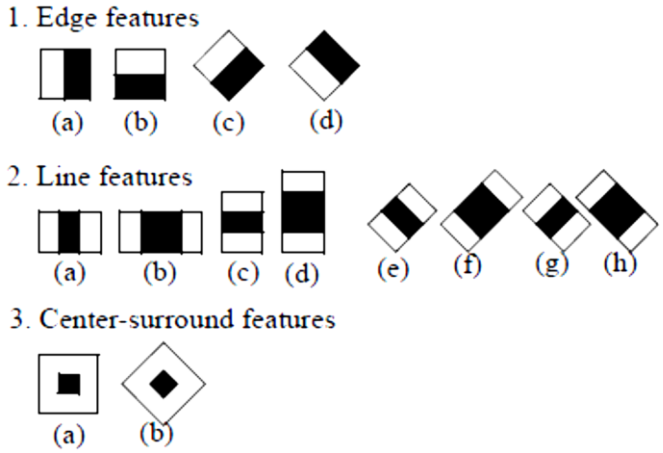
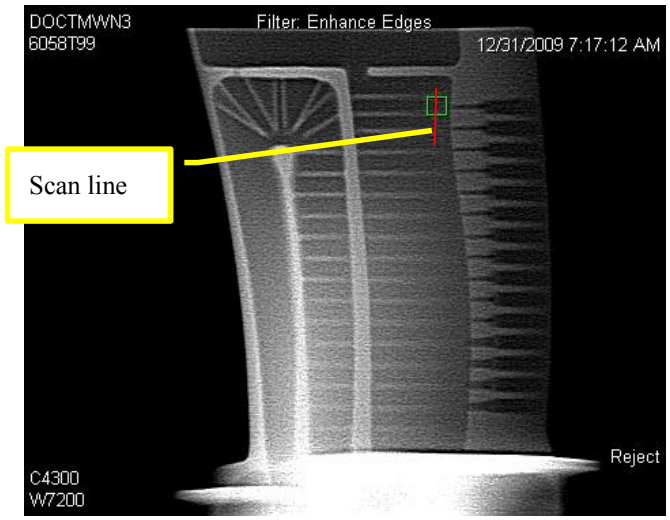


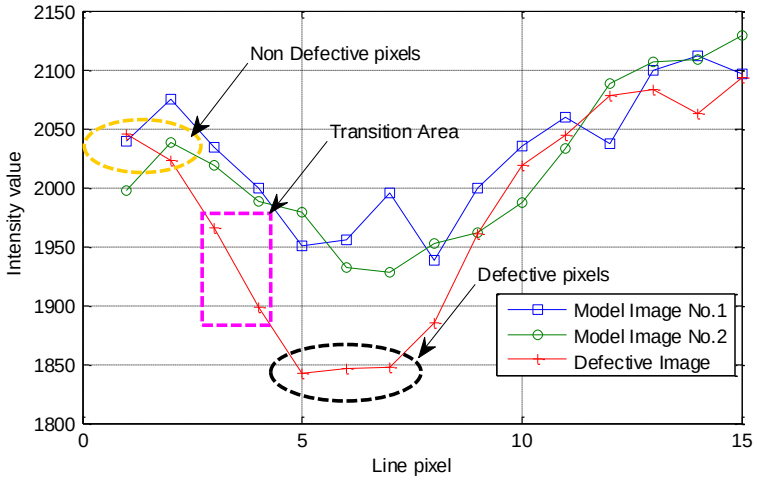
Figure 6.8. Extended Haar-like and center-surround features.

6.4.3 Modified Haar-like Feature

For turbine blades, the original or extended Haar-like features might not be suitable for defect recognition. Between a defect-free area and defective area, there might be a transition (fuzzy) area, that is, pixels in the transition area might be partly defective and partly non-defective. Including these pixels into feature calculation might make the distinction unobvious between the defect and non-defect. Figure 6.9 illustrates the transition area using an example of strong negative image undetected by ADR.



(a) Undetected image with strong negative defect



(b) Gray-level profiles of the scan line for the defective image in (a) and model images

Figure 6.9. An illustration of defective transition area.

Scanning through the green bounding box in Figure 6.9 (a), the gray-level profile of the scan line would go through a transition area (the purple bounding area), where the pixels' intensity of the defective image (the red profile) is slight lower than the pixels' intensity of the two model images (the blue profiles), but the intensity difference is not as obvious as the pixels in the black bounding eclipse between the defective image and model images. Including the pixels in the calculation of rectangle features would decrease the difference between defective area and non-defective area. Based on this observation, we propose a modified Haar-like feature, which is shown in Figures 6.10 (a) and (b).

(a) (b)

Figure 6.10. Proposed modified Haar-like feature. The sum of the pixels which lie within the black rectangles are subtracted from the sum of pixels in the yellow rectangles, with the pixels in the purple rectangle excluded in calculation.

The modified Haar-like features are represented as the following:

$$f_j = \text{Rect } Sum(\text{Y}) - \text{Rect } Sum(\text{B}),\ \text{j} = 3, 4, \ldots \tag{6.9}$$

where Y and B represent pixels in the Yellow and Black rectangle respectively, the length and width of the three rectangles are (Ly, Wy), (Lp, Wp) and (Lb, Wb) respectively.

We can view the purple rectangle as the guard band, which might be a region with pixels partly defective and is not largely different from either the Yellow rectangle or the Black rectangle.

6.5 Learning Decision Rules

Given the model images, the scan line and modified Haar-like features to be extracted at specific regions defined by human experts, our hypothesis, which is borne out by experiment, is that a normal region enclosing the feature space of the defect-free images can be defined by learning the feature space of the model images.

For each extracted scan line or modified Haar-like feature f_j of the image x, a threshold decision function $h_j(x)$ is learned from the model images to decide if the image x is defective or not. The threshold decision function $h_j(x)$ for a feature f_j is as the following:

$$h_j(x) = \begin{cases} 1 & \text{if } f_j(x) > \theta_j \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

where f_j is the j^{th} scan line or modified Haar-like feature, θ_j is the decision threshold. For a test image x, if there is a $f_j(x)$ greater than θ_j, then $h_j(x) = 1$, $\forall j = 1, 2, \ldots$, and the image is considered as defective.

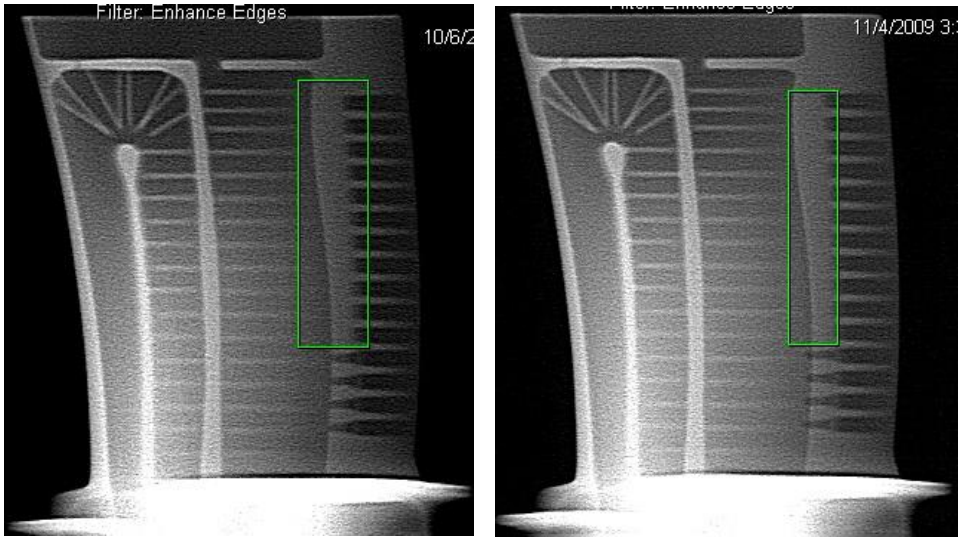
6.6 Experimental Results and Discussion

The scan line and modified Haar-like features are used for inspection of the undetected images from View 1 of blade type 'B' by ADR. For View 1 of blade type 'B', there are 9835 defect-free images, 173 strong negative defective images (S-), 62 strong positive defective images (S+), 3271 weak negative defective images and 99 weak positive defective images. The weak defective images correspond to blades with very minor anomalies and are not concerned about here as the corresponding blades can be used in production. We focus on the detection of strong defective images. A model set of 130 images is selected from the 9835 defect-free images using the ADR Model Optimizer.

The ADR system has a detection rate of 96.53% for S- set, and 48.39% for S+ set and a false alarm rate of 4.31%. 6 S- and 32 S+ images are not detected by ADR. The S+ is not as damaging as S- images. We pay the main attention to the detection of the S- images. Section 6.6.1 and 6.6.2 focus on the detection of the six undetected S- images using the scan line and modified Haar-like features respectively. Section 6.6.3 discusses the use of modified Haar-like features to replace the scan-line features in detection. Section 6.6.4 verifies the effectiveness of modified Haar-like features for identifying the detected S- images by ADR. For the completeness of the study, detection of positive defective images are investigated in Section 6.6.5.

6.6.1 Results of the Scan Line Feature Based Classifier

Figure 6.11 shows the two undetected S- images with defects labeled by the green bounding box. Scan line features are used to detect the two images.



(a) (b)

Figure 6.11. Two undetected defective images by ADR with defects labeled by the green bounding box.

Figure 6.12 shows the scan line feature space of the model images. From Figure 6.12, we see two decision rules h1(x) and h2(x) are learned from the model images based on scan line features f1 and f2. To be specific,

$$h_1(x) = \begin{cases} 1 & if\ f_1(x) > 0.46 \\ 0 & otherwise \end{cases}, \text{ and } \quad h_2(x) = \begin{cases} 1 & if\ f_2(x) > 0.27 \\ 0 & otherwise \end{cases}.$$

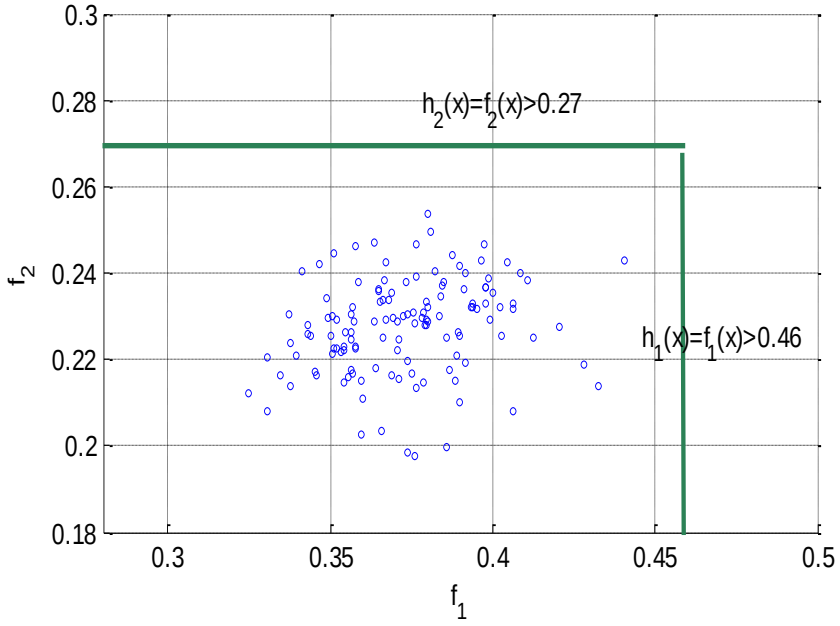


Figure 6.12. Decision rule learning based on the model set. The blue dots represents the feature points of the model images based on the scan line features f1 and f2.

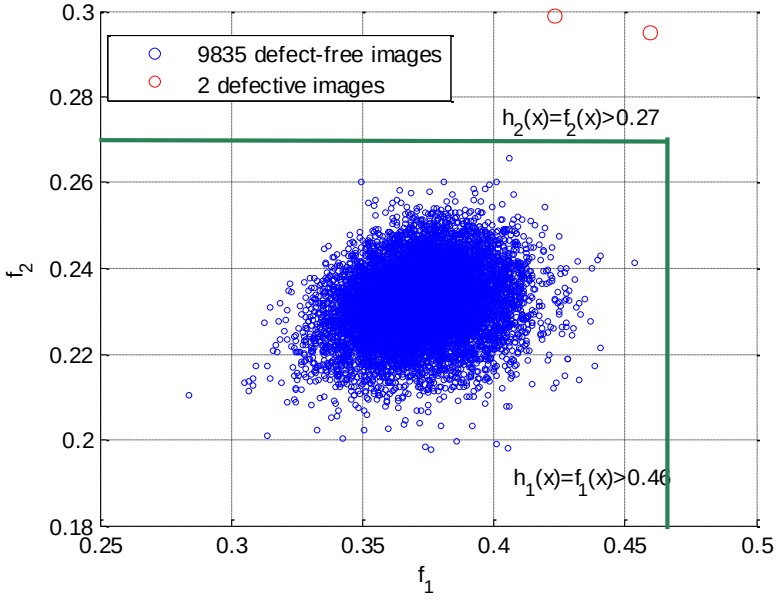


Figure 6.13. Distribution of the defect-free images and two undetected defective images by ADR.

Figure 6.13 shows the testing results using the two rules for the 9835 defect-free images and the two undetected defective images by ADR in Figure 6.11.

Seen from Figure 6.13, using the two decision rules formed from learning the model images, it is evident that the two S- defective images can be detected. All the 9835 defect-free images are correctly classified as non-defective as the distributions are within the learned normal feature space.

Figure 6.14 shows the feature distribution of the strong negative images with no defects in the labeled bounding box area in Figure 6.11. There are total 171 such images. From Figure 6.14, we see the decision ruled learned from the model images can correctly classify these images as non-defective in this area.

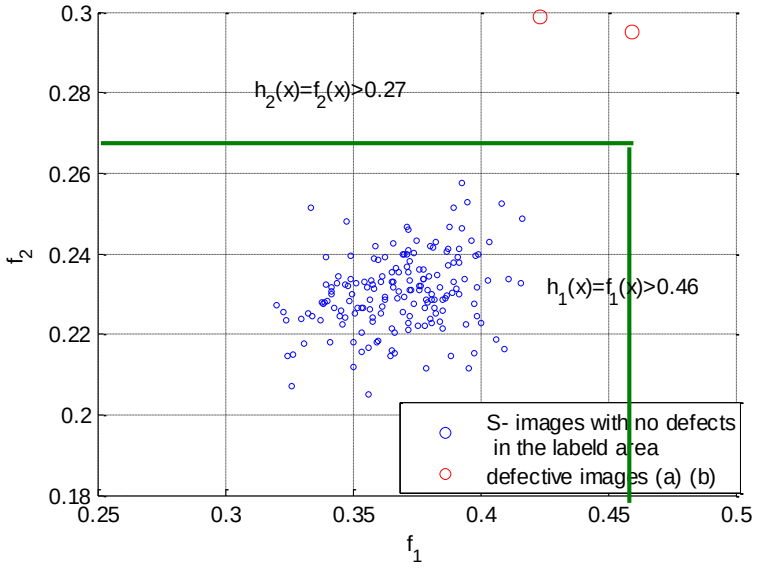
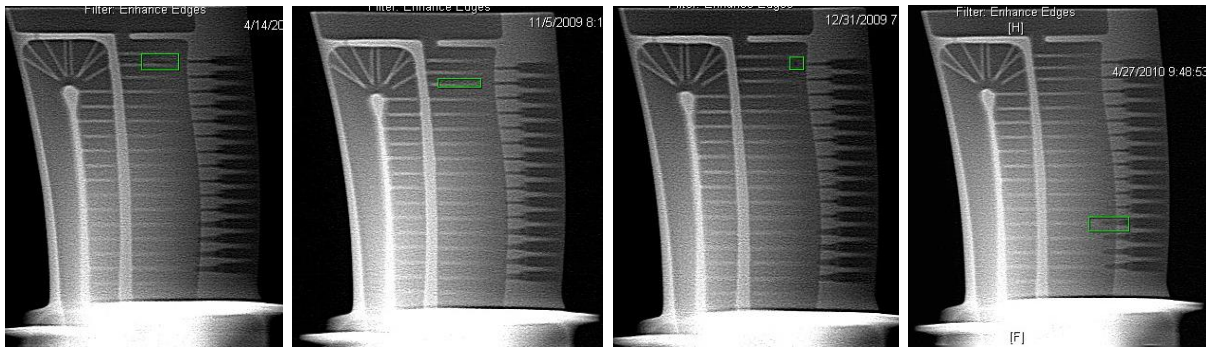


Figure 6.14. Decision rule learning based on the model set using the scan line features and testing results for the strong defective images with no defects in the labeled bounding box area in Figure 6.11.

6.6.2 Results of the Modified Haar-like Feature Based Classifier

Figure 6.15 shows the four undetected S- images with defects labeled by the green bounding box. Modified Haar-like features are used to detect these images.

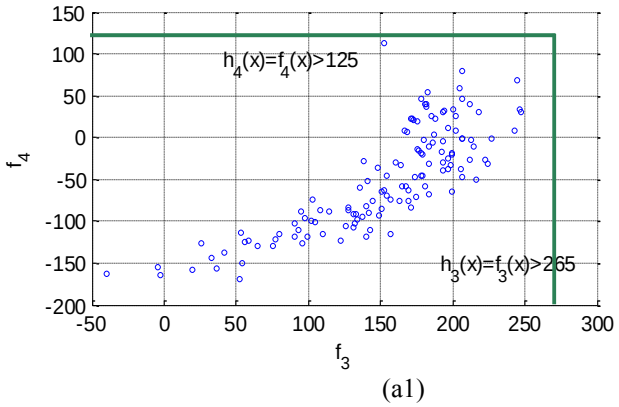


(a) (b) (c) (d)

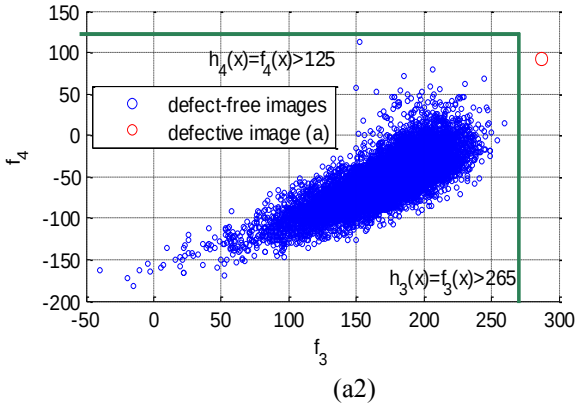
Figure 6.15. Four undetected S- images with defects labeled in the green bounding box.

Figures 6.16 (a1), (b1), (c1) and (d1) show the modified the distribution of the model images in the Haar-like feature space and the learned rules. Figures 6.16 (a2), (b2), (c2), and (d2) show the corresponding testing results using the learned rules for the 9835 defect-free images and the four undetected S- defective images by ADR respectively.

From Figures 6.16 (a1), (b1), (c1) and (d1), we see decision rules h3(x), h4(x)… h10(x) are learned from the model images based on the modified Harr-like features f3… f10. From Figures 6.16 (a2), (b2), (c2), and (d2), we see using the learned decision rules, the four defective images can be detected, and the distribution of all the 9835 defect-free images are within the normal feature space defined by the decision rules.



(a1)



(a2)

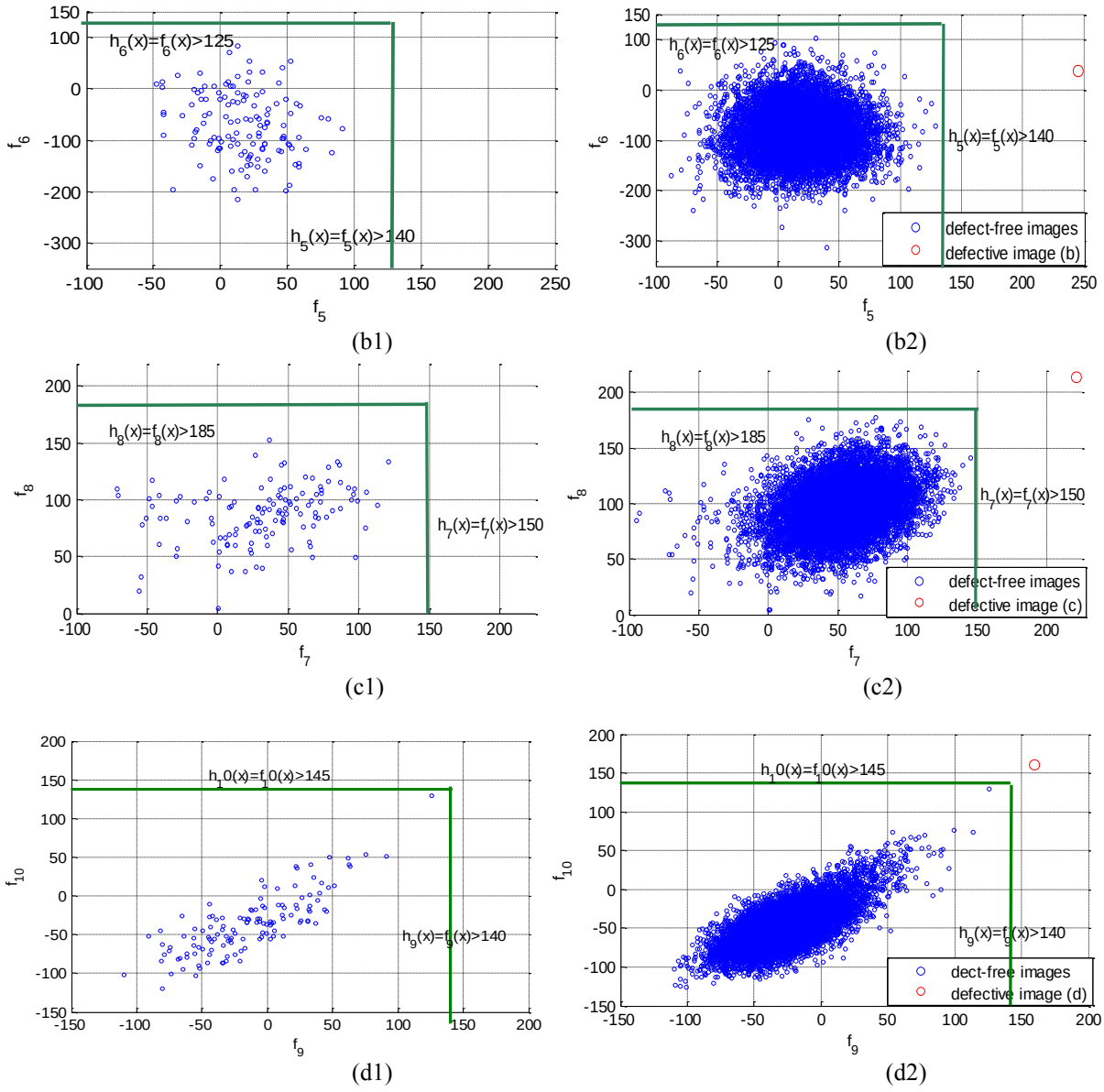
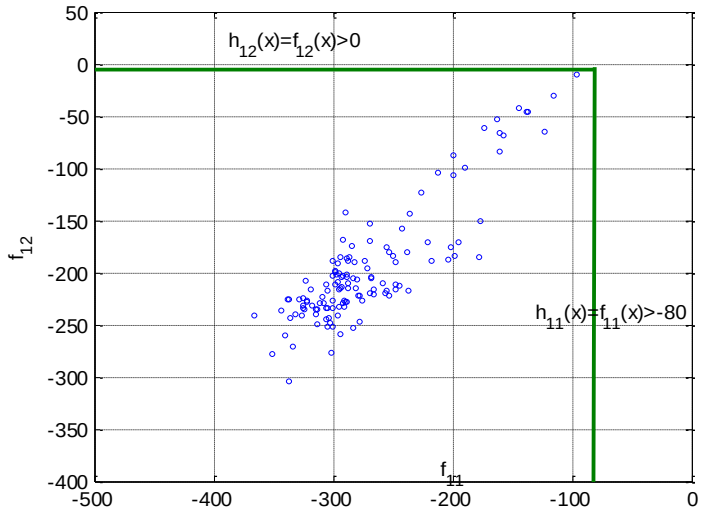


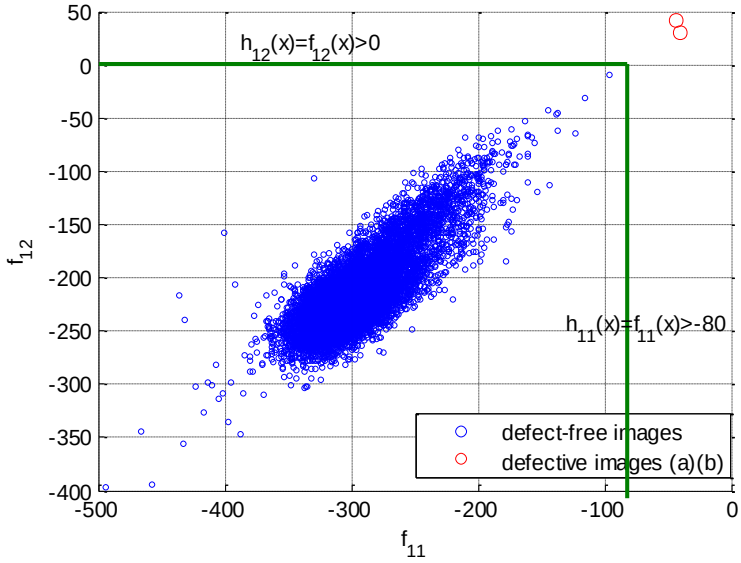
Figure 6.16. Decision rule learning based on the model set and testing results for all the defect-free image set and the undetected S- defective images shown in Figure 6.15.

6.6.3 Scan Line Feature Vs Modified Haar-like Feature

We investigate the possibility of using the modified Haar-like features to replace the scan line features, which were used for the detection of the undetected images in Figures 6.11 (a) and (b). Figure 6.17 shows the results using the modified Haar-like instead of the scan line features for decision rule learning and testing results of the decision rules for classifying the 9835 defect-free images and the defective images.



(a)



(b)

Figure 6.17. Decision rule learning based on the model set using the modified Haar-like features and testing results for the all the defect-free image set and the undetected defective images in Figure 6.11.

Seen from Figures 6.17 (a) and (b), using the two decision rules formed from learning model images, the two defective images can also be detected, and all the 9835 defect-free images in the feature space are within the normal region. Thus the modified Haar-like features can be used to replace the scan line features.

We check the decision rules for classification of the strong negative images with no defects in the labeled bounding box area in Figure 6.11. Figure 6.18 shows the distribution of these images in the modified Haar-like feature space. From Figure 6.18, we see these images are within the normal feature space in the labeled area.

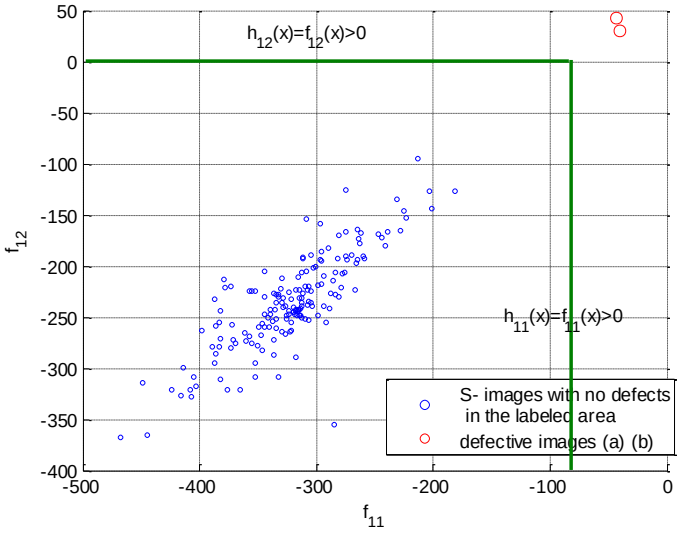
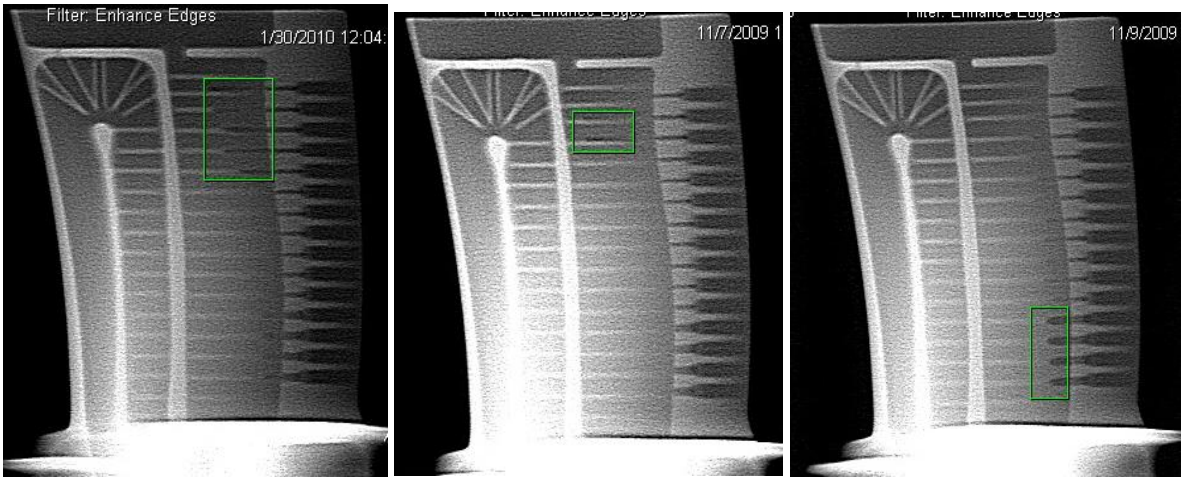


Figure 6.18. Decision rule learning based on the model set using the modified Haar-like features and testing results for the strong defective images with no defects in the labeled bounding box area in Figure 6.11.

6.6.4 Results of Using Modified Haar-like Features for Detected Strong Negative Images

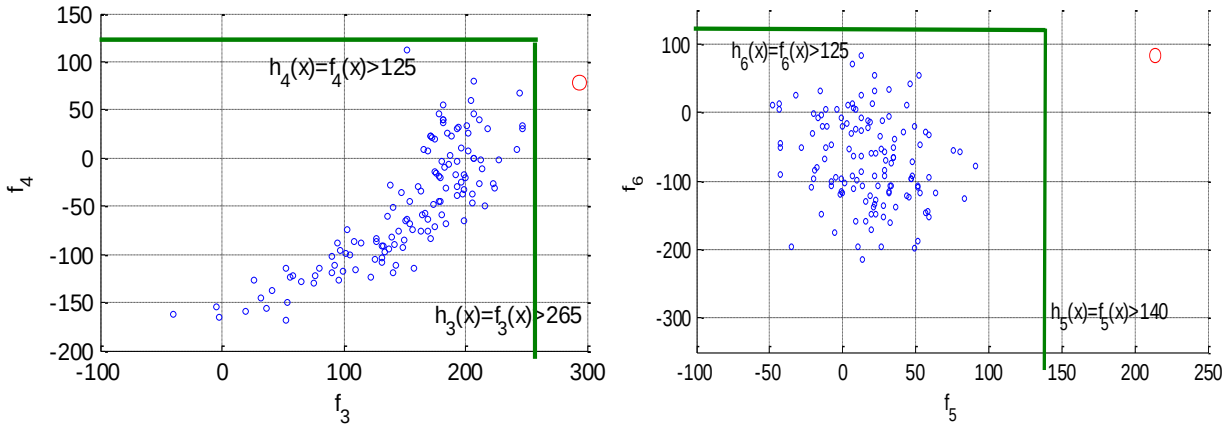
In previous sections, we see the modified Haar-like features based classifier can identify the undetected S-images by ADR. For the generality, we check if these features based classifier work for the detected S-images by ADR. Figures 6.19 (a), (b) and (c) list three such sample images.



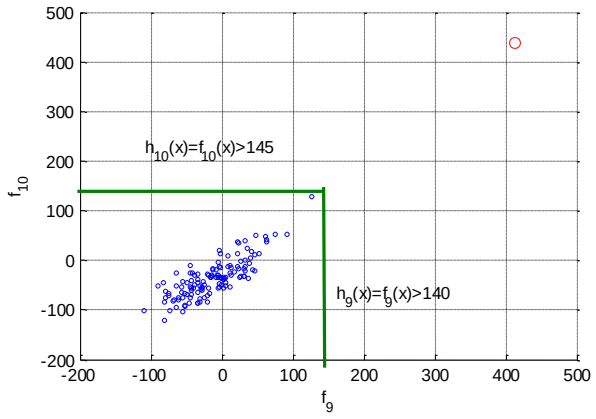
(a) (b) (c)

Figure 6.19. Sample detected images by ADR with strong negative defects labeled in the green bounding box.

The images in Figures 6.19 (a), (b) and (c) have defects indications that coincide with the images in Figures 6.15 (a), (b) and (d) respectively. It is expected that these images can be classified as defective using the learned decision rules in Figures 6.16 (a), (b) and (d) respectively. Figure 6.20 shows the distribution of the three images in the extracted corresponding feature space.



(a) (b)



(c)

Figure 6.20. The distribution of the images in Figure 6.19 (a), (b) and (c) in the corresponding modified Haar-like feature space. The red dot represents the images in Figures 6.19 (a), (b) and (c) respectively.

From Figures 6.20 (a), (b) and (c), we see the decision rules can correctly classify all the three images as defective in the labeled areas. We compare Figure 6.20 (c) and Figure 6.16 (d2), and can see that the features f_9 and f_{10} of the image in Figure 6.19 (c) are more far from the normal feature region than the image in Figure 6.15 (d). This matches the fact that the image in Figure 6.19 (c) has worse negative defects than the image in Figure 6.15 (d) in the labeled green bounding box area.

The modified Haar-like features based classifier can identify those detected S- images, which proves that the modified Haar-like features are able to characterize strong negative defects.

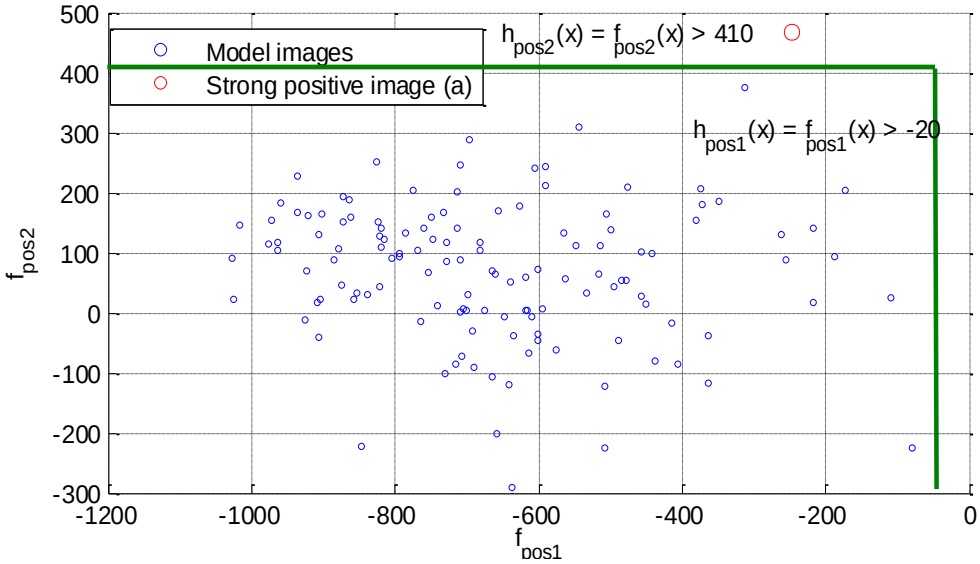
6.6.5 Results of Using Modified Haar-like Features for Strong Positive Images

The strong positive defective images corresponds to images of turbine blades with strong positive material defects. The strong positive defects are less damaging than the strong negative defects. For the completeness of the study, we investigate if the modified Haar-like features can be used to identify the strong positive defects in this section.

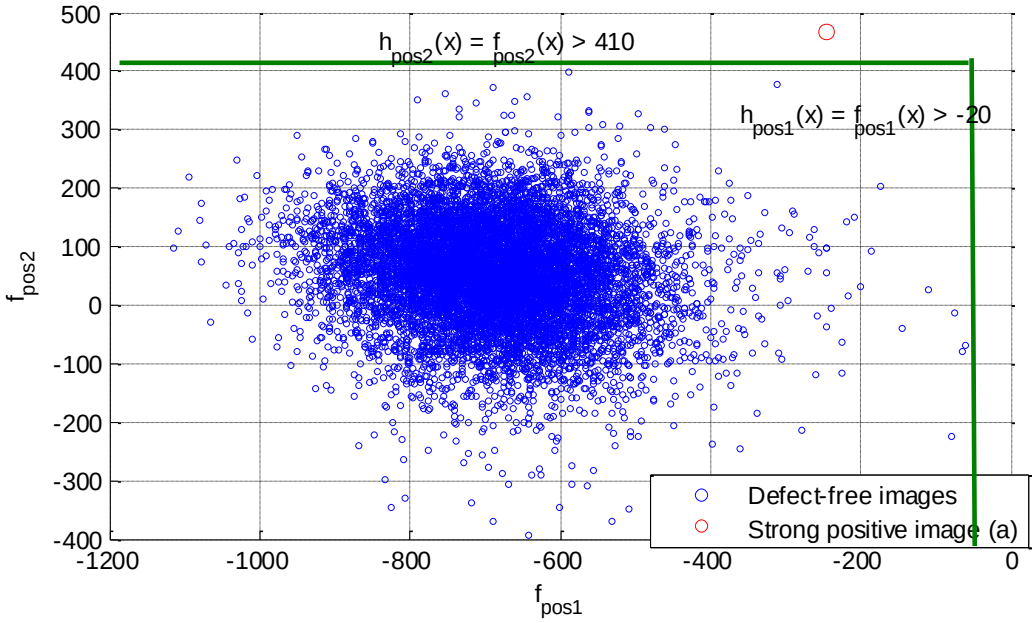
The positive defect indications are not obvious as the negative defects. Figures 6.21 (a), (b) and (c) list three images with strong positive defect indications. The images (a) and (b) are considered defective by ADR and can be called out, and the image (c) could not be detected by ADR. Figures 6.22 (a1) and (b1) show the decision rules learned from the model set based on the modified Haar-like features. Figures 6.22 (a2) and (b2) show the testing results for the defect-free image set and the strong positive defective images (a) and (b) in Figure 6.21.

(a) (b) (c)

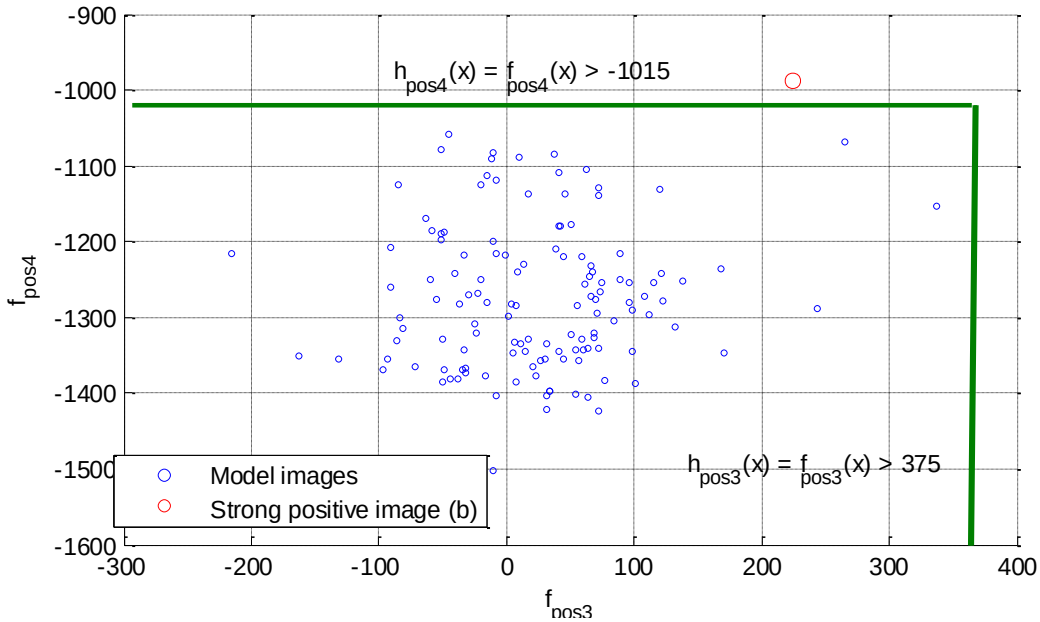
Figure 6.21. Sample images with strong positive defects labeled in the green bounding box.



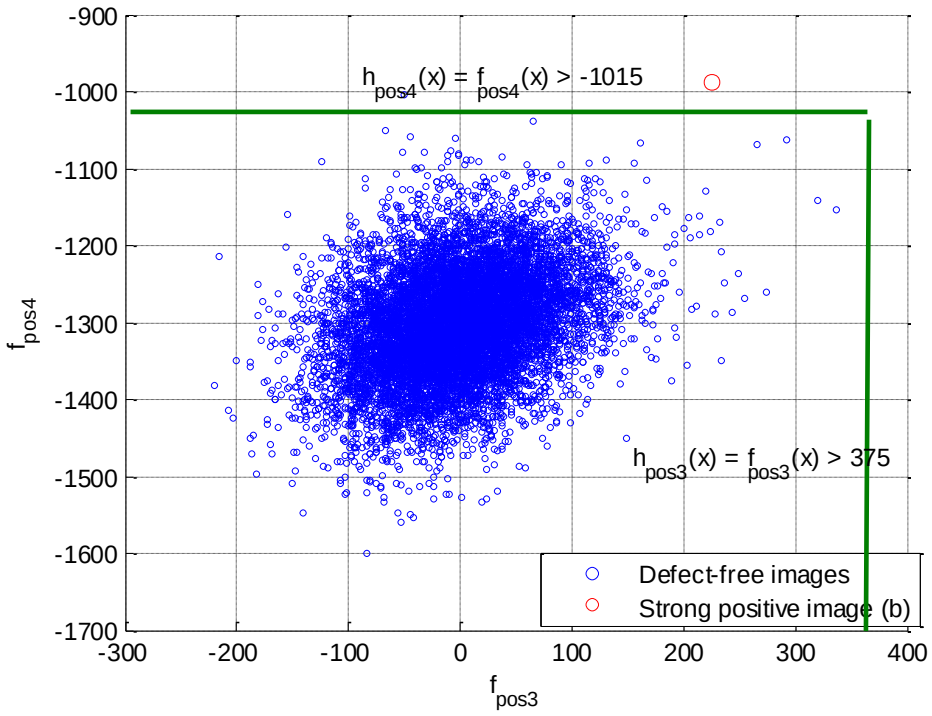
(a1)


h_pos2(x) = f_pos2(x) > 410
h_pos1(x) = f_pos1(x) > -20
Defect-free images
Strong positive image (a)
f_pos2
f_pos1
500
400
300
200
100
0
-100
-200
-300
-400
-1200
-1000
-800
-600
-400
-200
0

(a2)


h_pos4(x) = f_pos4(x) > -1015
h_pos3(x) = f_pos3(x) > 375
Model images
Strong positive image (b)
f_pos4
f_pos3
-900
-1000
-1100
-1200
-1300
-1400
-1500
-1600
-300
-200
-100
0
100
200
300
400

(b1)



(b2)

Figure 6.22. Decision rule learning based on the model set and testing results for the defect-free image set and the strong positive defective images in Figures 6.21 (a) and (b).

From Figures 6.22 (a1) and (b1), we see based on the modified Haar-like features decision rules can be learned to correctly identify the strong positive images (a) and (b) in Figure 6.21. The learned decision rules can also well separate the defect-free images from the strong positive images as shown in Figures 6.22 (a2) and (b2).

For the strong positive image in Figure 6.21 (c) undetected by ADR, decision rules are hardly to obtain due to the difficulty of the modified Haar-like features to separate it from the model set. More investigation is needed to find effective features that characterize this kind of positive defects. We see the positive defect in Figure 6.21 (c) has an oblique orientation, which are different from the horizontal or vertical oriented defects in Figures 6.21 (a), (b) and Figures 6.19 (a), (b) and (c). Figures 6.23 shows another image with oblique positive defects. Oblique defects are difficult to detect using the modified Haar-like features.

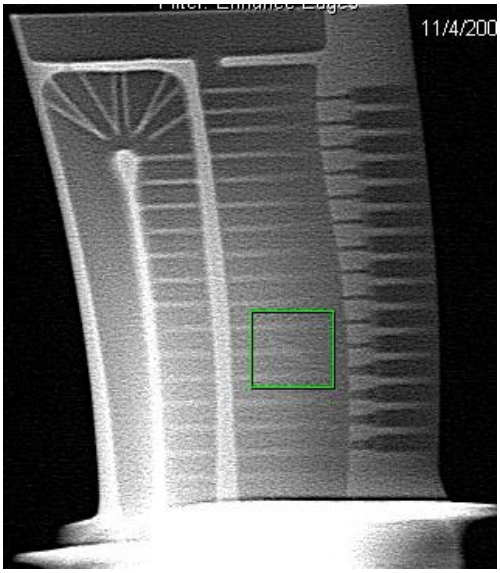
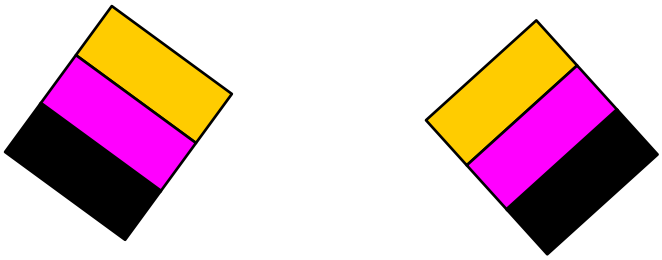


Figure 6.23. An image with oblique positive defects labeled in the green bounding box.

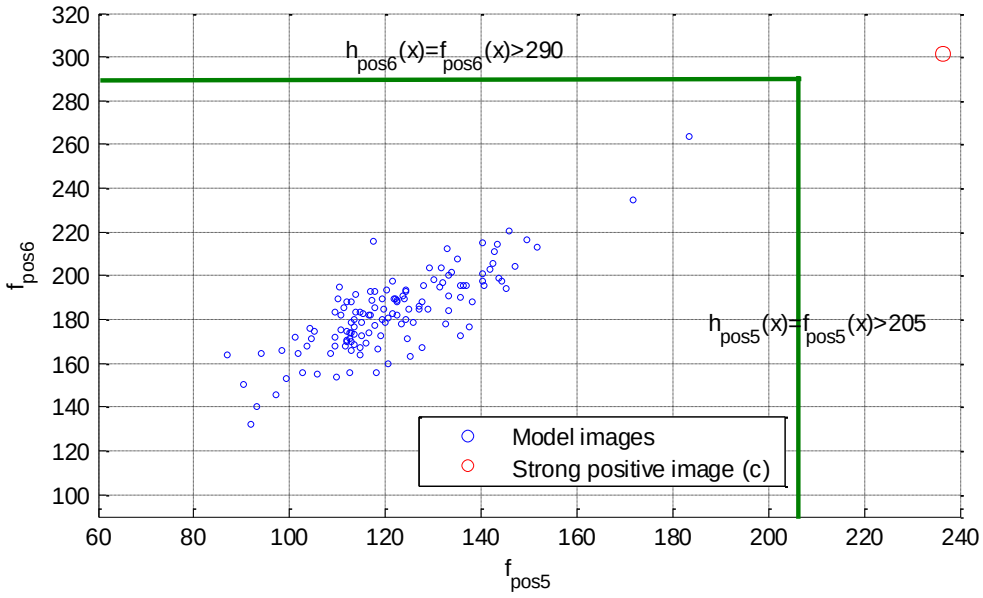
To fit the oblique defects, the modified Haar-like features are extended with oblique orientations as illustrated in Figures 6.24 (a) and (b). We expect that the extended Haar-like features be effective to discriminate the oblique defects in Figure 6.21(c) and Figures 6.23 from the good (defect-free) images.



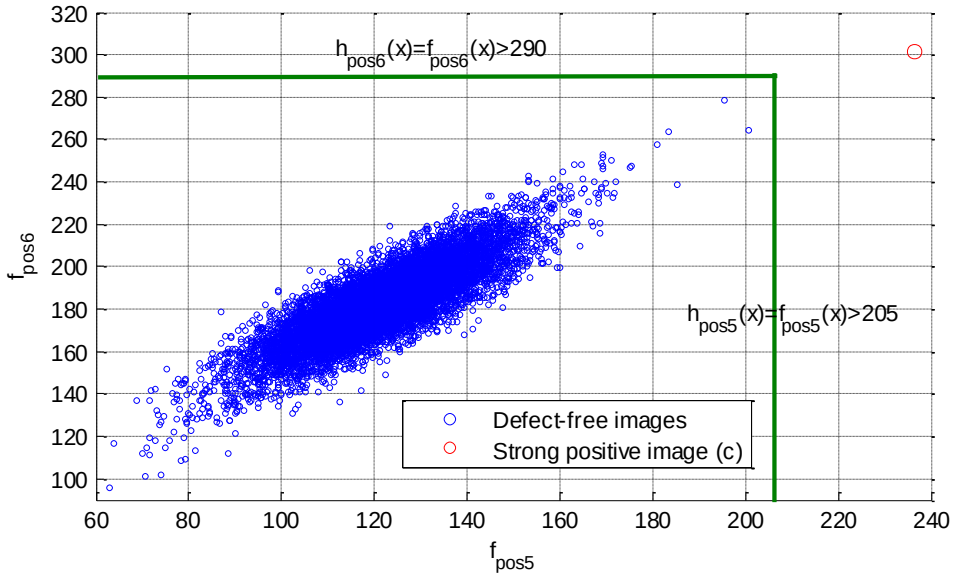
(a) (b)

Figure 6.24. Extended modified Haar-like feature. The sum of the pixels which lie within the black rectangle are subtracted from the sum of pixels in the yellow rectangle, with the pixels in the purple rectangle excluded in calculation.

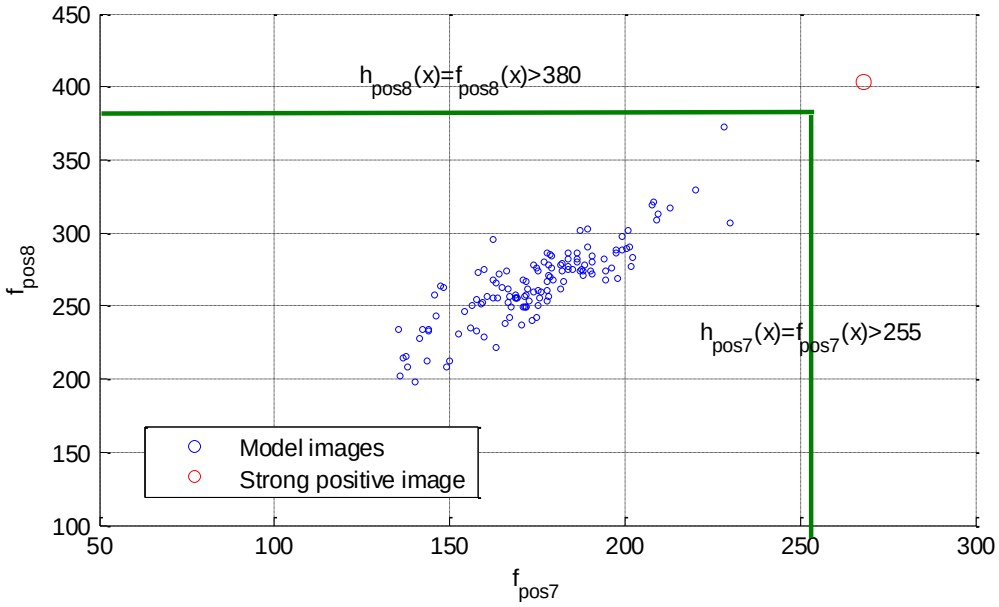
Figures 6.25 (a1) and (b1) show the decision rules learned from the model set based on the extended modified Haar-like features. Figures 6.25 (a2) and (b2) show the testing results for the detection of the images in Figure 6.21 (c) and Figures 6.23 and all the defect-free images using the decision rules.



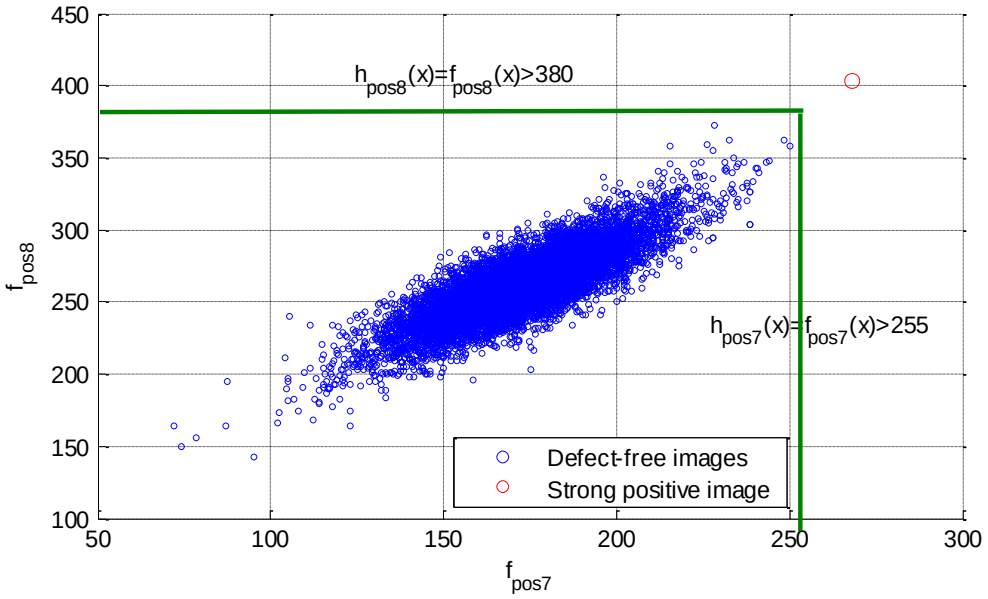
(a1)



(b1)



(a2)

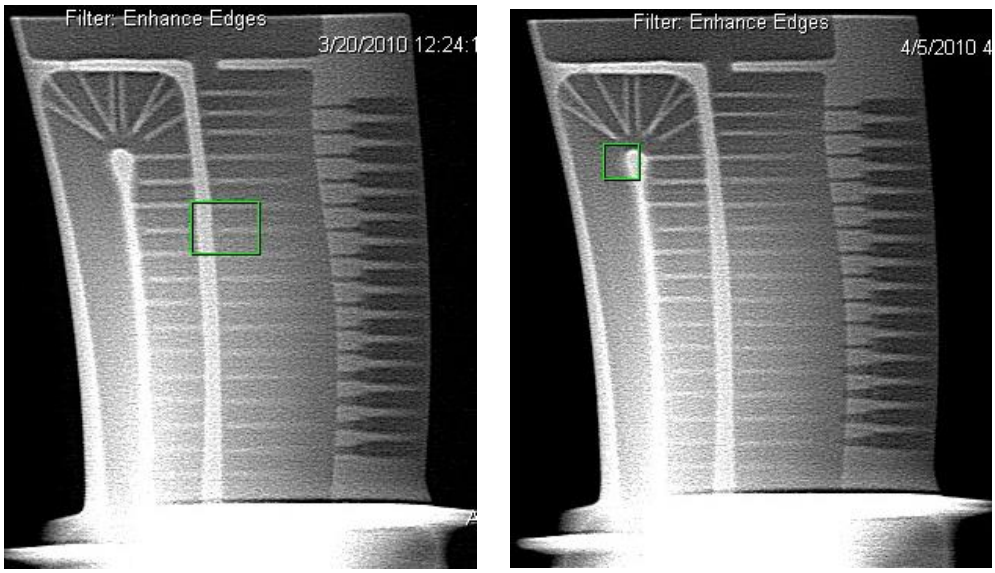


(b2)

Figure 6.25. Decision rule learned based on the model set and testing results for all the good defect-free images and the strong positive defective images in Figure 6.21 (c) and Figure 6.23.

From Figures 6.25 (a1) and (a2), we see the decision rules learned from the model set can correctly identify the strong positive images in Figure 6.21 (c) and Figure 6.23. From Figures 6.25 (b1) and (b2), we see the decision rules also well separate all the defect-free images from the defective images. This proves the effectiveness of the extended modified Haar-like features to detect the oblique positive defects.

The positive defects might occur anywhere in the turbine blades as discussed in Chapter 5. It might be difficult to use modified or extended modified Haar-like features to detect all of the strong positive defective images. Figures 6.26 (a) and (b) show two failed cases.



(a) (b)

Figure 6.26. Strong positive defective images failed to detect using the modified and extended modified Haar-like features with the defect labeled with green bounding box.

Defects in Images (a) and (b) in Figure 6.26 are not obvious and could not be detected by ADR. Based on the modified or extended modified Haar-like features, it is also hard to find decision rules to discriminate the two images from the defect-free images. More studies are needed to find features that can identify images with similar defect indications.

6.7 Conclusions

A new classifier based on scan line and modified Haar-like features extracted at defined regions in the image is proposed. The classifier learned decision rules defining the normal feature space from the model images, and based on the decision rules images are classified as defective or defect-free. Experimental results show that the classifier can detect all the undetected strong negative images by ADR. The detection rate has been increased from 96.53% to 100% without increasing any false alarm rate. From the results, we see the modified Haar-like features can be used to replace the scan line features. As expected, the modified Haar-like features based classifier is also able to detect the detected strong negative images by ADR. The modified Haar-like features are extended with oblique orientations. Using the modified and extended modified Haar-like features, the classifier can detect some strong positive defective images.

7 Conclusions

This dissertation addresses the challenges involved in the inspection of industrial parts based on radiographic images. Approaches and methods were developed to facilitate and improve an assisted defect recognition (ADR) system for turbine blades of jet engines. Main contributions were the following.

- An automatic approach – the Model Optimizer was proposed to select a reference model set from a large set of defect-free images for ADR. Though extensive experiments, we found that the Model Optimizer could select a small model set yielding a low false alarm rate with acceptable detection rate. Based on the selected model set, the ADR can identify the defect indication types and location relatively accurately. The Model Optimizer outperforms than manual approach and has been successfully applied in the production line of GE Aviation.
- An adaptive model selection procedure was developed for ADR to adapt parts' variations in the production process based on the Model Optimizer. The procedure could detect significant variations automatically and update the reference model image set. The reference set is adapted to represent the current production process, leading to a low false alarm rate with acceptable detection rate continuously. The developed procedures involves little human intervention in the selection process, and can adapt a model set for ADR to identify the defect indication types and location relative accurately.
- A systematic methodology for evaluating the impact of defective images in the model set on the performance of the ADR was proposed. The methodology uses McNemar's test for measure the performance difference for model sets with and without defective images. The number of defective images that can be tolerated in the model set for each type of defects was determined. Testing results show that strong negative defective images are more damaging than the positive defective images if to be included into the model set.
- A new classifier based on scan-line and modified Haar-like features was put forward by combining ADR to improve defect detection. The classifier learns decision rules defining the

normal feature space based on the model set at specific regions. The classifier can detect not only all the undetected strong negative images by ADR, but also those detected strong negative images and some strong positive images. Further investigation proved that the modified Haar-like features can replace the scan line features for detection. The modified Haar-like features are effective for the detection of horizontal and vertical defects. We extended the modified Haar-like features for identifying defects with oblique orientations. Some positive defective images were failed to detect using the modified and extended modified Haar-like features. This will be possible part of the future study. More extensive validations are needed in order to deploy the proposed classifier in production.

Bibliography

[1] M. Carrasco and D. Merry, Automatic multiple view inspection using geometrical tracking and feature analysis in aluminum wheels, Machine Vision and Applications, vol.22, no.1, pp. 157-170, 2011.

[2] S. M. Tam and K. C. Cheung, Genetic algorithm based defect identification system, Expert Systems with Applications, vol.18, no.1, pp.17-25, 2000.

[3] D. Mery and D. Filbert, Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence, IEEE Trans. Robot. Autom., vol.18, no.6, pp. 890-901, 2002.

[4] Z. Sun, R. Kaucic, P. Mendoca and A. Can, A Statistical Approach to Industrial Anomaly Detection, in Proc. of the 10th European conference on Nondestructive testing, 2010.

[5] D. Mery, T. Jaeger and D. Filbert, A review of methods for automated recognition of casting defects, Insight, vol.44, no.7, pp.428-436, 2002.

[6] F. Zhao, P. R. S. Mendonca and R. Kaucic, Image-Based Automated Defect Recognition via Statistical Learning of Minkowski Functionals, in Proc. of the 10th European conference on Nondestructive testing, 2010

[7] L. Fillatre, I. Nikiforov and F. Retraint, A Simple Algorithm for Defect Detection From a Few Radiographies, J. of Computers., vol.2, no.6, pp.26-34, 2007.

[8] H. Boerner and H. Strecker, Automated X-ray Inspection of Aluminium Casting, IEEE Trans. Pattrn. Anal. Machine Intell., vol.10, no.1, pp. 79-81, 1988.

[9] S. Lawson and G. Parker, Intelligent Segmentation of Industrial Radiographic Images Using Neural Networks, in Proc. of SPIE, Mach. Vision Applicat. Syst. Integration III, vol.2347, pp. 245-255, Nov. 1994.

[10] T. Wenzel and R. Hanke, Fast Image Processing on Die Castings, in Proc. Anglo-German Conf. NDT Imaging and Signal Processing, Oxford, U.K., Mar. 1998.

[11] A. Kehoe and G. A. Parker, An Intelligent Knowledge Based Approach for the Automated Radiographic Inspection of Castings, NDT & E International, vol.25, no.1, pp. 23-36, 1992.

[12] A. Vincent, V. Rebuffel, R. Guillemaud, L. Gerfault and P. Coulon, Defect Detection in Industrial Casting Components Using Digital X-ray Radiography, Insight, vol.44, no.10, pp. 632-327, 2004.

[13] M. Sofia and D. Redouane, Shapes Recognition System Applied to the Non Destructive Testing in Proc. of the 8th European Conference on Non-Destructive Testing, 2002.

[14] J. Kosanetzky and H. Putzbach, Modern X-ray Inspection in the Automotive Industry, in Proc. of the 14th World Conference on Non-Destructive Testing, 1996.

[15] X. Xiao, J. Quan, A. Ferro, C. Han, X. Zhou and W. Wee, On selecting reference image models for anomaly detection in industrial systems, In Proc. SPIE 8856, Applications of Digital Image Processing XXXVI, 88560O, San Diego, California, August 2013

[16] T. Stocker and T. Wenzel, Automatic X-ray Inspection with Dynamic Reference Data Sets, International Symposium on Digital Industrial Radiology and Computed Tomography, Berlin, Germany, Jun. 2011.

[17] A. Bifet and R. Gavaldà, Learning from Time-Changing Data with Adaptive Windowing, In Proceedings of the SIAM International Conference on Data Mining, pp. 443–448, Minneapolis, MN, 2007.

[18] L. Fillatre, I. Nikiforov, and F.Retraint, ε-optimal non-bayesian anomaly detection for parametric tomography. IEEE Transactions on Image Processing, vol.17, no.11, pp.1985–1999, 2008.

[19] T. Stocker, and T. Wenzel. Automatic X-ray Inspection with Dynamic Reference Data Sets, International Symposium on Digital Industrial Radiology and Computed Tomography, Berlin, Germany, June 2011.

[20] L. E. Bryant, Nondestructive Testing Handbook, 2nd ed. Columbus, OH: Amer. Soc. Nondestructive Testing, vol. 3, Radiography &Testing, 1985.

[21] G. Wang and T. W. Liao, Automatic identification of different types of welding defects in radiographic images, NDT&E Int., vol. 35, no.8, pp. 519–528, 2002.

[22] P. Baniukiewicz, Automated Defect Recognition and Identification in Digital Radiography, Journal of Nondestructive Evaluation, vol.33, no.3, pp. 327-334, Sep. 2014.

[23] F. Herold, and R. Grigat, A New Analysis and Classification Method for Automatic Defect Recognition in X-Ray Images of Castings, 8th ECNDT, Barcelona, Jun. 2002.

[24] A. Stone, What the Devil is Fully Automatic Real-Time X-Ray Inspection, CSNDT Journal, vol.21, No.1, pp.6, 8-10, 12, 2000.

[25] F. Herold, K. Bavendiek, and R. Grigat, A third generation automatic defect recognition system, In Proceedings of the 16th World Conference on Non-Destructive Testing (WCNDT 2004), Montreal, 2004.

[26] J. M. Kosanetzky and H. Putzbach, Modern x-ray inspection in the automotive industry, In Proc. 14th World Conference of NDT. New Delhi, Dec. 1996.

[27] YXLON. The new image of automatic defect recognition. Technical Articles of YXLON International, 2002. http://www.yxlon.com/technical articles.htm.

[28] D. Mery, Automated radioscopic inspection of aluminum die castings, Materials Evaluation, vol.65, no.6, pp. 643–647, 2006.

[29] J. Demˇsar, Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, vol.7, pp.1–30, 2006.

[30] T. G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, vol. 10, pp.1895–1923, 1998.

[31] S. García, F. Herrera, An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons, Journal of Machine Learning Research, vol. 9, pp. 2677-2694, 2008.

[32] R. R. Bouckaert and E. Frank, Evaluating the replicability of significance tests for comparing learning algorithms. In D. Honghua, R. Srikant, and C. Zhang, editors, Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings. Springer, 2004.

[33] P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, vol.30, pp.1145–1159, 1997.

[34] P. H. Westfall, J. F. Troendle and G. Pennello, Multiple McNemar test. Biometrics, vol. 66, no.4, pp.1185–1191, 2010.

[35] A. Walker and J. Shostak. Common statistical methods for clinical research with SAS examples 3rd edition, SAS Institute, 2010.

[36] D. J. Sheskin. Handbook of parametric and nonparametric statistical procedures, Chapman & Hall, 2007.

[37] L. T. Warren, D. Li and Y. Li, Detection of welding flaws from radiographic images with fuzzy clustering methods," Fuzzy Sets and Systems vol.108, no.2, pp. 145-158, 1999.

[38] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in Proceedings of ICASSP, 1989.

[39] P. Armitage. Statistical Methods in Medical Research. Blackwell Scientific Publications, 1971.

[40] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, vol.12, pp.153–157, 1947.

[41] A. K. Jain, Statistical Pattern Recognition: A Review, IEEE Tran. on Pattern Analysis and Machine Intelligence, vol.22, no 1, Jan. 2000.

[42] R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, 2nd edition, John Wiley $ Sons, Inc., 2001.

[43] Graves, Mark & Bruce G. Batchelor, Machine Vision for the Inspection of natural products, Springer P 5, 2003.

[44] D. Tsai, C. Chang and S. Chao, Micro-crack inspection in heterogeneously textured solar wafers using anisotropic diffusion, Image and Vision Computing, vol.28, no.3, pp.491-501, 2010.

[45] P. Viola and M. Jones, Robust real-time face detection, International Journal of Computer Vision, vol. 57, no.2, 2004.

[46] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, In IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[47] X. Xiao , P. Howard, A. Ferro, T. Ma, C. Y. Han, X. Zhou and W. Wee, Automated X-ray Inspection with Imperfect Reference Data, Preprint, 2015.

[48] L. Chmielewski, M. Nieniewski, M. Skłodowski and W. Cudny, Detection of surface defects and irregularities of ferrites. In M. Weigl, editor, in Proc. 31st Polish Solid Mechanics Conference SolMec'96 - Book of Abstracts, pp. 63-64, Mierki, Poland, Sep. 1996.

[49] Z. Sun and A.Hoogs, Image comparison by compound disjoint information. In IEEE International Conference on Computer Vision and Pattern Recognition, pp. 857–862, New York, NY, Jun. 2006.

[50] J. Derrac, S. García, D. Molina, and F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation, vol. 1, no. 1, pp. 3–18, 2011.

[51] J. Higgins, Introduction to Modern Nonparametric Statistics, Duxbury Press, 2003.

[52] J. Luengo, S. García and F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and Interpretability, Soft Computing, vol.13, no.10, pp. 959-977, 2009.

[53] X. Xiao, A. Ferro, T. Ma, C. Y. Han, X. Zhou, and W. Wee, Adaptive Reference Image Set Selection in Automated X-Ray Inspection," Journal of Electrical and Computer Engineering, vol. 2014, pp 1-7, 2014. doi:10.1155/2014/794526.

[54] M. P. Boyce, Gas Turbine Engineering Handbook (3rd edition), Oxford, Elsevier, 2006

[55] B. Trawiński, M. Smętek, Z. Telec and T. Lasota, Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms, Int. J. Appl. Math. Comput. Sci. vol.22, no.4, pp. 867-881, 2012.

[56] Zar, J., Biostatistical Analysis, 5th Edn., Prentice Hall, Upper Saddle River, NJ, 2009.

[57] J. Luengo, S. García and F. Herrera, A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests, Expert Systems with Applications, vol.36, pp.7798–7808, 2009.

[58] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics, vol.1, pp. 80–83, 1945.

[59] P. H. Westfall, J. F. Troendle and G. Pennello, Multiple McNemar Tests. Biometrics, vol.66, pp.1185–1191, 2010.

[60] W. Leisenring, T. Alonzo and M.S. Pepe, Comparisons of predictive values of binary medical diagnostic tests for paired designs, Biometrics, vol. 56, pp.345–351, 2000.

[61] V. L. Durkalski, Y. Y. Palesch, S. R. Lipsitz, F. Philip and P. F. Rust, The analysis of clustered matched-pair data. Statistics in Medicine, vol.22, pp. 2417–2428, 2003.

[62] R. H. Lyles, J. M. Williamson, H. M. Lin and C. M. Heilig, Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. Biometrics. vol.61, pp. 287–294, 2005.

[63] R. Szeliski, Image Alignment and Stitching: A Tutorial. Foundations and Trends in Computer Graphics and Computer Vision, vol.2, pp. 1-104, 2006.

[64] Z. Sun, A. Can, J. Janning, R. Kaucic, P. Mendonca, and J. Portaz, Method and system for identifying defects in NDT image data. U.S. Patent 8131107 B2, filed May 12, 2008, and issued March 6, 2012.

[65] R. Lienhart and J. Maydt, An Extended Set of Haar-like Features for Rapid Object Detection, IEEE ICIP 2002, vol. 1, pp. 900-903, Sep. 2002.